An International ICT Journal
Featuring Industry–University–Institute Cooperation and
Indexed in Scopus

ISSN 1673-5188
CN 34-1294/TN

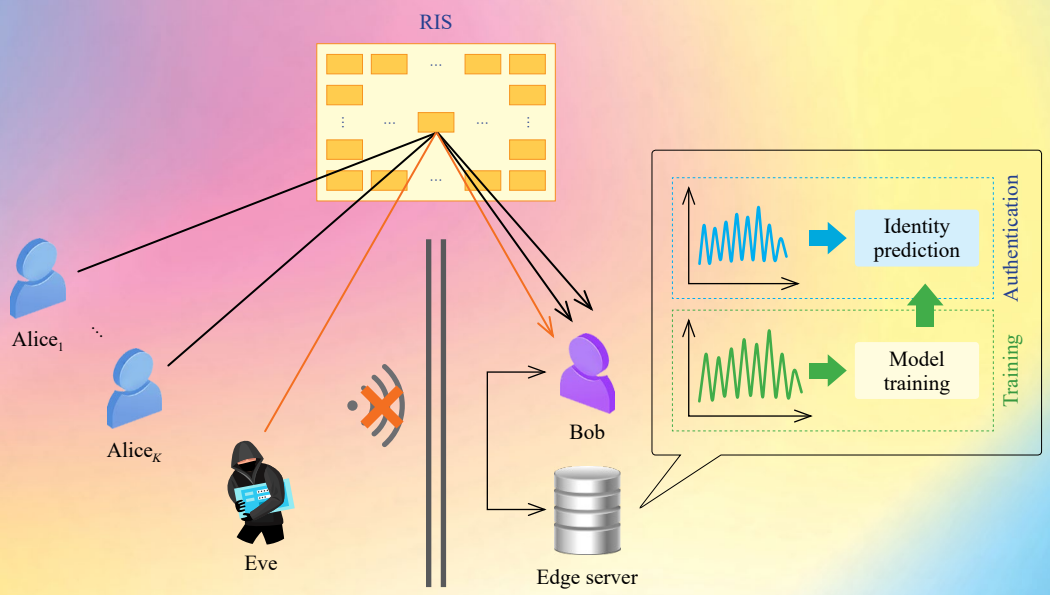
ZTE COMMUNICATIONS

中兴通讯技术(英文版)

http://zte.magtechjournal.com

March 2025, Vol. 23 No. 1

Special Topic:
Native Intelligence at the Physical Layer



(See Fig. 1 on P. 22 and Fig. 2 on P. 23)

ISSN 1673-5188
03

The 10th Editorial Board of ZTE Communications

CONTENTS

ZTE COMMUNICATIONS
March 2025 Vol. 23 No. 1 (Issue 90)

Special Topic ▶ Native Intelligence at the Physical Layer

Research Papers ▶

Serial parameters: CN 34–1294/TN*2003*q*16*122*en*P*¥30.00*2200*15*2025-03

ZTE Communications Guidelines for Authors

Remit of Journal

ZTE Communications publishes original theoretical papers, research findings, and surveys on a broad range of communications topics, including communications and information system design, optical fiber and electro-optical engineering, microwave technology, radio wave propagation, antenna engineering, electromagnetics, signal and image processing, and power engineering. The journal is designed to be an integrated forum for university academics and industry researchers from around the world.

Manuscript Preparation

Manuscripts must be typed in English and submitted electronically in MS Word (or compatible) format. The word length is approximately 3 000 to 8 000, and no more than 8 figures or tables should be included. Authors are requested to submit mathematical material and graphics in an editable format.

Abstract and Keywords

Each manuscript must include an abstract of approximately 150 words written as a single paragraph. The abstract should not include mathematics or references and should not be repeated verbatim in the introduction. The abstract should be a self-contained overview of the aims, methods, experimental results, and significance of research outlined in the paper. Three to eight carefully chosen keywords must be provided with the abstract.

References

Manuscripts must be referenced at a level that conforms to international academic standards. All references must be numbered sequentially intext and listed in corresponding order at the end of the paper. References that are not cited in-text should not be included in the reference list. References must be complete and formatted according to *ZTE Communications* Editorial Style. A minimum of 10 references should be provided. Footnotes should be avoided or kept to a minimum.

Copyright and Declaration

Authors are responsible for obtaining permission to reproduce any material for which they do not hold copyright. Permission to reproduce any part of this publication for commercial use must be obtained in advance from the editorial office of *ZTE Communications*. Authors agree that a) the manuscript is a product of research conducted by themselves and the stated co-authors; b) the manuscript has not been published elsewhere in its submitted form; c) the manuscript is not currently being considered for publication elsewhere. If the paper is an adaptation of a speech or presentation, acknowledgement of this is required within the paper. The number of co-authors should not exceed five.

Content and Structure

ZTE Communications seeks to publish original content that may build on existing literature in any field of communications. Authors should not dedicate a disproportionate amount of a paper to fundamental background, historical overviews, or chronologies that may be sufficiently dealt with by references. Authors are also requested to avoid the overuse of bullet points when structuring papers. The conclusion should include a commentary on the significance/future implications of the research as well as an overview of the material presented.

Peer Review and Editing

All manuscripts will be subject to a two-stage anonymous peer review as well as copyediting, and formatting. Authors may be asked to revise parts of a manuscript prior to publication.

Biographical Information

All authors are requested to provide a brief biography (approx. 100 words) that includes email address, educational background, career experience, research interests, awards, and publications.

Acknowledgements and Funding

A manuscript based on funded research must clearly state the program name, funding body, and grant number. Individuals who contributed to the manuscript should be acknowledged in a brief statement.

Address for Submission

http://mc03.manuscriptcentral.com/ztecom

Guest Editorial >>>

Special Topic on Native Intelligence at the Physical Layer

Guest Editors

YANG Kun

JIN Shi

XIANG Luping

The rapid evolution of wireless communication technologies, particularly with the advent of 5G and the impending transition to 6G, has underscored the need for innovative strategies to enhance the performance, reliability, and efficiency of communication systems. One such promising approach gaining significant attention is the concept of native intelligence at the physical layer (PHY). This forward-thinking concept integrates advanced algorithms and AI-driven solutions directly into the physical layer, transforming the way communication systems are managed and optimized in real time. The incorporation of native intelligence at the physical layer offers tremendous potential to meet the growing demands of future communication networks. By embedding artificial intelligence (AI) algorithms into the physical layer, these intelligent systems can autonomously adapt to dynamic channel conditions, thereby improving spectral efficiency, enhancing error correction, and ensuring robust communication even in highly challenging and fluctuating environments. Native intelligence is poised to become a crucial enabler for the advanced features promised by 6G networks, such as ultra-reliable low-latency communication, massive connectivity, and intelligent wireless ecosystems.

In this special issue, we aim to spotlight the latest advancement and research development in the field of native intelligence at the physical layer. We have invited high-quality submissions that explore the theoretical underpinnings and innovative use cases of AI in enhancing the physical layer of communication systems. The call for papers has garnered a series of excellent submissions, reflecting the growing interest and momentum in this emerging area. Following two rigorous rounds of peer review, the following seven papers are presented. These papers cover topics such as spatio-temporal channel state information (CSI) prediction for massive multiple input multiple output (MIMO), reconfigurable intelligent surfaces (RIS) -enhanced communication security, AI-based physical-layer authentication for 6G, rethinking source-channel coding, AI-native networks for 6G optimization, device activity detection in massive MIMO, and efficient primary synchronization signal (PSS) detection using convolutional neural networks (CNN), each demonstrating significant advancements in performance, security, and efficiency. The papers are organized as follows.

The first paper, titled "Efficient Spatio-Temporal Predictive Learning for Massive MIMO CSI Prediction", introduces a novel spatio-temporal predictive network (STPNet) that improves CSI prediction in massive MIMO systems. The STPNet model integrates both CSI feedback and prediction modules using deep learning techniques to capture spatio-temporal correlations. This approach improves the accuracy of CSI prediction, especially in scenarios with high mobility or feedback delays, outperforming traditional methods under various channel conditions.

The second paper, titled "RIS-Enabled Simultaneous Transmission and Key Generation with PPO: Exploring Security Boundary of RIS Phase Shift", investigates the use of RIS to enhance both communication security and transmission efficiency. The paper presents an integrated communication and security (ICAS) design that combines simultaneous transmis-

DOI:10.12142/ZTECOM.202501001
Citation (Format 1): YANG K, JIN S, XIANG L P. Native intelligence at the physical layer [J]. ZTE Communications, 2025, 23(1): 1–2. DOI: 10.12142/ZTECOM.202501001
Citation (Format 2): K. Yang, S. Jin, L. P. Xiang, "Native intelligence at the physical layer," *ZTE Communications*, vol. 23, no. 1, pp. 1–2, Mar. 2025. doi: 10.12142/ZTECOM.202501001.

sion and key generation (STAG). By optimizing RIS phase shifts through a proximal policy optimization (PPO) algorithm. The proposed system significantly improves security and convergence stability, demonstrating a 90% performance improvement in "one-time pad" communication compared with traditional methods.

The third paper, titled "Endogenous Security Through AI-Driven Physical-Layer Authentication for Future 6G Networks", explores the use of AI to enhance physical-layer security for 6G networks. The paper focuses on physical-layer authentication (PLA), leveraging the unique randomness and space-time-frequency characteristics of the wireless channel to provide secure identity signatures. The authors propose a graph neural network (GNN)-based PLA method that outperforms traditional authentication schemes in terms of accuracy, addressing emerging security challenges in 6G networks.

The fourth paper, titled "Separate Source Channel Coding Is Still What You Need: An LLM-Based Rethinking", challenges the conventional joint source channel coding (JSCC) paradigm and advocates for separate source channel coding (SSCC). The authors propose leveraging large language models (LLMs) for source coding and error correction code transformers (ECCT) for channel coding, showing that SSCC offers superior performance over JSCC. The paper provides an in-depth analysis of the compatibility challenges between semantic communication approaches and digital communication systems, demonstrating the efficiency of SSCC in modern communication contexts.

The fifth paper, titled "Exploration of NWDAF Development Architecture for 6G AI-Native Networks", explores the role of AI-native networks in 6G, focusing on the network data analytics function (NWDAF). The paper proposes a novel architecture that integrates real-time data collection, model training, and AI-driven decision-making to optimize network resource management. Through a vertical scaling use case on Kubernetes, the authors demonstrate the practical benefits of AI in improving network management and resource allocation, with the XGBoost model showing superior predictive performance.

The sixth paper, titled "Device Activity Detection and Channel Estimation Using Score-Based Generative Models in Massive MIMO", addresses the challenge of joint device activity detection and channel estimation in massive MIMO systems. The authors propose a score-based generative model for robust channel estimation, which adapts well to the complex and dynamic environments typical of massive MIMO systems. Simulation results show exceptional precision in channel estimation, with errors reduced to as low as −45 dB, and demonstrate high accuracy in detecting active devices. This method significantly improves the performance of network resource allocation and device activity detection in large-scale systems.

The seventh paper, titled "Efficient PSS Detection Algorithm Aided by CNN", proposes a fast PSS detection algorithm based on the correlation characteristics of PSS time-domain superposition signals. By incorporating CNN to estimate frequency offsets, the paper addresses potential accuracy issues caused by these offsets during the PSS detection process. The proposed method reduces computational complexity and improves communication speed, with simulation results demonstrating its effectiveness in enhancing PSS detection efficiency.

To conclude, we hope this special issue on native intelligence at the physical layer serves as a significant step forward in integrating intelligent algorithms directly into the physical layer of communication systems. Finally, we sincerely express our gratitude to all the authors and reviewers for their invaluable contributions, and we trust that the insights and innovations presented will inspire new directions for research and development in this exciting and evolving field.

Biographies

YANG Kun received his PhD from the Department of Electronic & Electrical Engineering, University College London (UCL), UK. He is currently a Chair Professor in the School of Intelligent Software and Engineering, Nanjing University, China. He is also an affiliated professor of University of Essex, UK. His main research interests include wireless networks and communications, communication-computing cooperation, and new AI for the wireless. He has published 500+ papers and filed 50 patents. He serves on the editorial boards of a number of IEEE journals (e.g., *IEEE WCM*, *TVT*, and *TNB*). He is a deputy editor-in-chief of *IET Smart Cities Journal*. He has been a judge of GSMA GLOMO Award at MWC Barcelona since 2019. He was a recipient of 2024 IET Achievement Medal and the 2024 IEEE CommSoft TC Technical Achievement Award. He is a member of Academia Europaea (MAE), a Fellow of IEEE, a Fellow of IET and a Distinguished Member of ACM.

JIN Shi received his BS degree in communication engineering from Guilin University of Electronic Technology, China in 1996, MS degree from Nanjing University of Posts and Telecommunications, China in 2003, and PhD degree in information and communications engineering from the Southeast University, China in 2007. From June 2007 to October 2009, he was a research fellow with the Adastral Park Research Campus, University College London, UK. He is currently with the faculty of the National Mobile Communications Research Laboratory, Southeast University. His research interests include wireless communications, random matrix theory, and information theory. He is serving as an area editor for the *Transactions on Communications* and *IET Electronics Letters*. He was an associate editor for the *IEEE Transactions on Wireless Communications*, *IEEE Communications Letters*, and *IET Communications*. Prof. JIN and his co-authors have been awarded the 2011 IEEE Communications Society Stephen O. Rice Prize Paper Award in the field of communication theory, the 2024 IEEE Communications Society Marconi Prize Paper Award, the IEEE Vehicular Technology Society 2023 Jack Neubauer Memorial Award, a 2022 Best Paper Award, and a 2010 Young Author Best Paper Award by the IEEE Signal Processing Society.

XIANG Luping received his BE degree (Hons.) from Xiamen University, China in 2015, and PhD degree from the University of Southampton, UK in 2020. From 2020 to 2021, he was a research fellow with the Next Generation Wireless Group, University of Southampton. In November 2021, he joined the University of Electronic Science and Technology of China (UESTC) and in September 2024, he joined Nanjing University, China as an assistant professor. In 2024, he was honored with the Xiaomi Young Scholar Award, and also co-founded the company Accelercomm. His research interests include native intelligence at wireless communication, end-to-end transmission technology, computer vision, and integrated sensing and communication transmission.

Efficient Spatio-Temporal Predictive Learning for Massive MIMO CSI Prediction

CHENG Jiaming[1], CHEN Wei[1], LI Lun[2,3], AI Bo[1]

(1. School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China;
2. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China;
3. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202501002

https://kns.cnki.net/kcms/detail/34.1294.TN.20250303.1371.004.html, published online March 4, 2025

Manuscript received: 2025-02-23

Abstract: Accurate channel state information (CSI) is crucial for 6G wireless communication systems to accommodate the growing demands of mobile broadband services. In massive multiple-input multiple-output (MIMO) systems, traditional CSI feedback approaches face challenges such as performance degradation due to feedback delay and channel aging caused by user mobility. To address these issues, we propose a novel spatio-temporal predictive network (STPNet) that jointly integrates CSI feedback and prediction modules. STPNet employs stacked Inception modules to learn the spatial correlation and temporal evolution of CSI, which captures both the local and the global spatio-temporal features. In addition, the signal-to-noise ratio (SNR) adaptive module is designed to adapt flexibly to diverse feedback channel conditions. Simulation results demonstrate that STPNet outperforms existing channel prediction methods under various channel conditions.

Keywords: massive MIMO; deep learning; CSI prediction; CSI feedback

Citation (Format 1): CHENG J M, CHEN W, LI L, et al. Efficient spatio-temporal predictive learning for massive MIMO CSI prediction [J]. ZTE Communications, 2025, 23(1): 3 - 10. DOI: 10.12142/ZTECOM.202501002
Citation (Format 2): J. M. Cheng, W. Chen, L. Li, et al., "Efficient spatio-temporal predictive learning for massive MIMO CSI prediction," *ZTE Communications*, vol. 23, no. 1, pp. 3 - 10, Mar. 2025. doi: 10.12142/ZTECOM.202501002.

1 Introduction

Future 6G communication systems are expected to support significantly higher demands from mobile broadband services[1]. As a representative 6G scenario, ultra-massive multiple-input multiple-output (MIMO) systems critically depend on real-time, accurate, and reliable channel state information (CSI)[2]. In frequency division duplex (FDD) systems, user equipment (UE) estimates downlink CSI and feeds it back to the base station (BS) via uplink transmission. However, the increasing number of antennas has dramatically expanded the feedback overhead, thereby placing a substantial burden on limited bandwidth resources. Recently, deep learning (DL) techniques have been introduced to compress CSI and reduce feedback overhead[3-4]. Specifically, DL-based CSI feedback utilizes an encoder to compress the CSI into codewords at the UE and a decoder at the BS to reconstruct the CSI from these codewords[5]. This approach has been demonstrated to outperform traditional codebook-based feedback methods in terms of effectiveness[6]. In Ref. [7], SwinCFNet is proposed for a CSI feedback task, which utilizes the Swin Transformer to extract long-range dependency information from CSI.

However, due to changes in the scattering environment and user mobility, the channel varies rapidly over time. In mobile scenarios, processing delay in the CSI feedback process makes the CSI received by the BS outdated, leading to a significant degradation in system performance. The authors in Ref. [8] theoretically analyze the impact of CSI delay on the channel. To mitigate the performance degradation caused by channel aging, accurate and timely CSI prediction becomes increasingly essential, which leverages the temporal correlation between historical CSI and future channel states. Besides, in recent years, digital twins have emerged as a revolutionary technology for visualizing, predicting, and analyzing the interactions between digital models and the physical world[9]. The design of digital twins relies on the virtual mapping of physical products, using real-time data and information from the field. High-precision time series prediction of wireless channel information in physical entities is crucial to building a digital twin environment[10].

Traditional methods for CSI prediction, such as the linear extrapolation model[11] and the autoregressive (AR) model[12], rely on statistical and mathematical formulations that struggle

This work was supported in part by the Natural Science Foundation of China under Grant Nos. U2468201 and 62221001 and ZTE Industry-University-Institute Cooperation Funds under Grant No. IA20240420002.

to capture the dynamic complexity of realistic wireless channels. In contrast, DL-based models, with their capacity for capturing nonlinear relationships and their flexibility in handling large datasets, offer a promising alternative. Inspired by the great potential of the recurrent neural networks (RNNs) and their variants in time series modeling, an RNN-based predictor[13] and a long short-term memory (LSTM)-based predictor[14] have been proposed. In Ref. [15], a transformer-based parallel channel prediction model is introduced to accurately predict time-varying channels, which avoids the error propagation problem in classical sequential prediction methods. Additionally, the authors in Ref. [16] propose a joint framework for channel feedback and prediction, leveraging the convolutional LSTM (ConvLSTM) to exploit temporal correlations. However, these existing methods primarily focus on the temporal correlation, while overlooking the array and frequency correlations crucial for further improvement.

In this paper, we propose a novel spatio-temporal predictive network (STPNet) for CSI prediction in massive MIMO systems. STPNet employs a joint CSI feedback and prediction framework, where the feedback network compresses and reconstructs CSI while capturing inter-antenna and inter-subcarrier correlations. The core prediction network consists of several cascaded Inception modules to learn the spatio-temporal features from the codewords by group convolutions. Using joint training, STPNet eliminates the error propagation issues found in separate module designs. Furthermore, we introduce a signal-to-noise ratio (SNR) adaptive module to dynamically adjust input tokens according to real-time SNR variations, enabling more robust adaptation to changing communication conditions. Numerical results show that STPNet outperforms other predictive methods across diverse channel scenarios.

2 System Model

We consider the downlink of an FDD massive MIMO system with $N_t \gg 1$ transmitting antennas at the BS and a single receiving antenna at the UE. The number of sub-carriers is N_c. The received signal at the n-th subcarrier can be expressed as:

$$y_n = \boldsymbol{h}_n^H \boldsymbol{v}_n x_n + z_n \quad (1),$$

where $\boldsymbol{h}_n \in \mathbb{C}^{N_t}$, $\boldsymbol{v}_n \in \mathbb{C}^{N_t}$, $x_n \in \mathbb{C}$, and $z_n \in \mathbb{C}$ denote the channel vector, the precoding vector, the transmitted data symbol and the additive noise at the n-th subcarrier, respectively. The downlink CSI can be denoted by:

$$\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_n]^{\mathrm{T}} \in N_c \times N_t \quad (2).$$

Since the elements of the channel matrix are complex numbers, the total number of CSI parameters is $2N_cN_t$. However, as the number of antennas in future massive MIMO systems grows, the size of the CSI matrix might exceed the uplink's feedback capacity.

To tackle the challenge of payload size reduction, we implement a framework that compresses the channel matrix $\boldsymbol{H}$ into a low-dimensional codeword $\boldsymbol{s}$ of size $M\times1$ at the UE, which can be formulated as:

$$\boldsymbol{s} = f_{\mathrm{en}}(\boldsymbol{H}; \theta_{\mathrm{en}}) \quad (3),$$

where $f_{\mathrm{en}}(\cdot)$ represents the function of the encoder and θ_{en} denotes its parameter. The compression ratio (CR) is defined as $\gamma = \frac{M}{2N_cN_t}$. Then, the encoded vector $\boldsymbol{s}$ is transmitted via a noisy channel. In our work, we consider the widely used additive white Gaussian noise (AWGN) channel. The channel output vector $\hat{\boldsymbol{s}}$ received by the BS is expressed as:

$$\hat{\boldsymbol{s}} = \eta(\boldsymbol{s}, \sigma) = \boldsymbol{s} + \boldsymbol{n} \quad (4),$$

where each component of the noise vector $\boldsymbol{n}$ is independently sampled from a Gaussian distribution, i.e., $\boldsymbol{n} \sim \mathcal{N}(0, \sigma^2\boldsymbol{I})$, and σ^2 is the noise power.

The structure of AI-based CSI feedback is illustrated in Fig. 1a. However, in high-speed mobile scenarios, the channel matrix varies rapidly over time. Due to the feedback delay and channel aging problems, directly feeding back the channel at the current slot leads to a mismatch between the feedback channel and the actual channel. To address this issue, a CSI prediction module is introduced at the BS. Our proposed AI-based joint CSI feedback and prediction framework, shown in Fig. 1b, performs prediction at the codeword level. Let $\hat{\boldsymbol{s}}^{(t)}$ and $\bar{\boldsymbol{s}}^{(t+1)}$ denote the codeword of the t-th slot and the predicted codeword of the $(t+1)$-th slot, respectively. We adopt the received historical codewords from the past P slots to simultane-

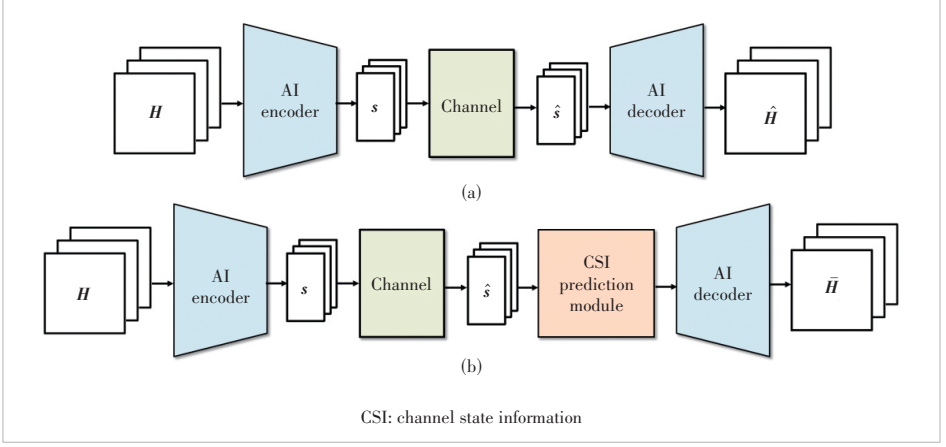


Figure 1. (a) Structure of AI-based CSI feedback; (b) Our proposed AI-based joint CSI feedback and prediction framework

ously predict the future codewords for the next L consecutive slots simultaneously, which can be expressed as:

$$(\bar{\boldsymbol{s}}^{(t+1)},\cdots,\bar{\boldsymbol{s}}^{(t+L)}) = f_{\text{pre}}(\hat{\boldsymbol{s}}^{(t-P+1)},\cdots,\hat{\boldsymbol{s}}^{(t)};\theta_{\text{pre}}) \quad (5),$$

where $f_{\text{pre}}(\cdot)$ represents the function of the prediction module and θ_{pre} denotes the corresponding parameter set. Subsequently, the BS reconstructs the channel matrix from the predicted future codewords as follows.

$$\bar{\boldsymbol{H}} = f_{\text{de}}(\bar{\boldsymbol{s}};\theta_{\text{de}}) \quad (6),$$

where $f_{\text{de}}(\cdot)$ represents the function of the decoder and θ_{de} denotes the parameter set of the decoder. $\bar{\boldsymbol{H}}$ is the recovered channel matrix.

3 Design of STPNet

3.1 Network Architecture

Compared with simple CSI feedback, joint CSI feedback and prediction can more effectively mitigate the distortion caused by feedback delays and channel aging. In a separate feedback and prediction architecture, each module is optimized and designed independently, so the local optimum of each component may not yield a globally optimal outcome. In contrast, the joint architecture employs end-to-end training to reduce error propagation between modules, resulting in more accurate predicted CSI.

Building on the advantages of the joint feedback and prediction architecture, we present an overview of our STPNet model in Fig. 2a. STPNet consists of a CSI encoder, SNR adaptive modules, a CSI prediction module and a CSI decoder. The encoder is used to compress the CSI into codewords and extract spatial features of the channel matrix at UE. The CSI prediction module, serving as the network's core, operates at the codeword level. The prediction module leverages the spatial and temporal correlation of historical channel characteristics to forecast future codewords. The SNR adaptive modules, integrated into both the encoder and decoder, dynamically modulate intermediate tokens based on instantaneous channel quality. Finally, the decoder aggregates and processes the predicted codewords to produce the final CSI output at the BS.

We employ the SwinCFNet architecture to implement the CSI encoder and decoder within STPNet. Built upon the Swin Transformer, SwinCFNet delivers superior performance in CSI feedback tasks. First, it effectively reduces feedback data while aggregating spatial-frequency domain CSI features to support the prediction module. Second, this design captures long-range dependencies, exploiting both inter-frequency and inter-antenna correlations within the channel matrix, ultimately enhancing the accuracy of the predicted output. Ref. [7] presents a detailed description of the SwinCFNet architecture.

In the core prediction module, an Inception architecture is introduced to learn the temporal evolution by capturing and updating spatio-temporal features, as shown in Fig. 2b. Motivated by Refs. [17] and [18], cascaded Inception blocks are

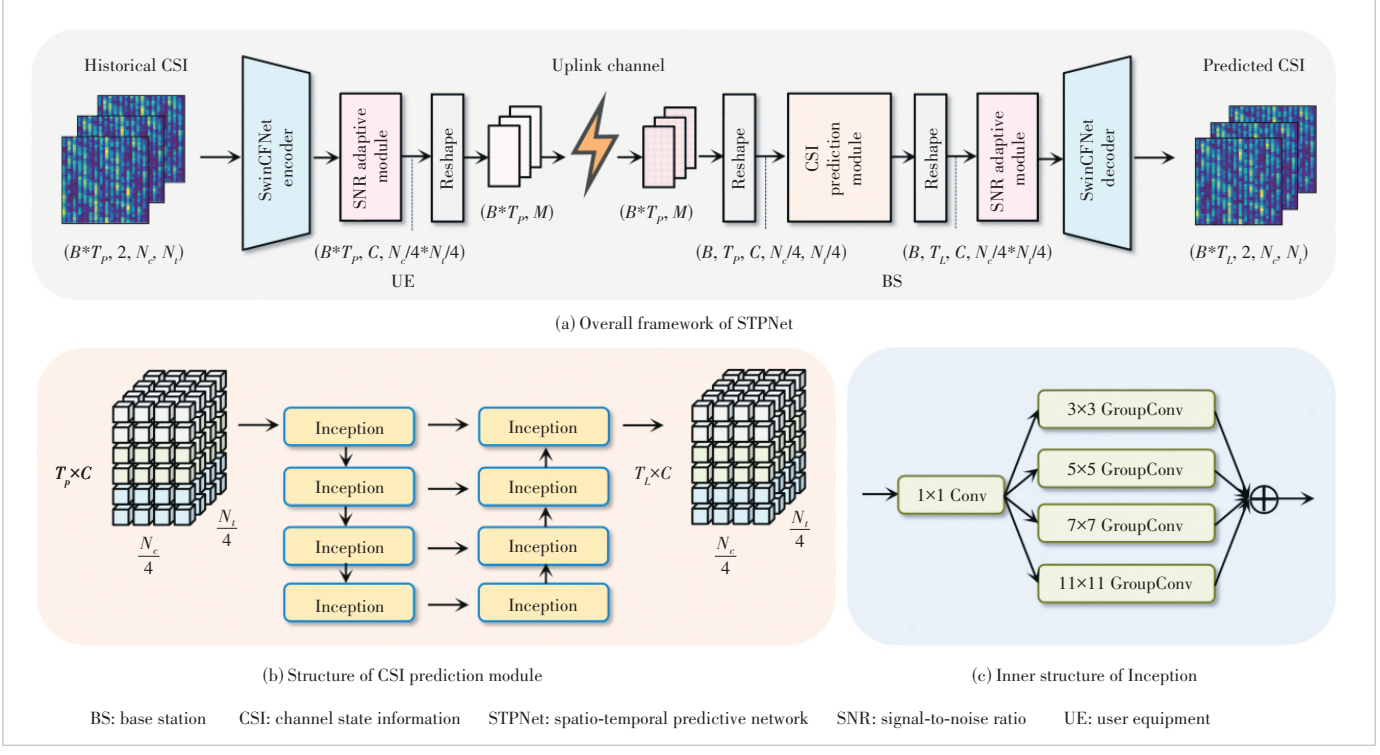


Figure 2. Network architecture of STPNet

employed. These blocks primarily consist of convolution layers with 1×1 kernels, followed by parallel GroupConv2D operations. The inner structure of Inception is illustrated in Fig. 2c. Here, the 1×1 Conv2D layer is used to increase the depth of the network and enhance representational capacity. To extract diverse local patterns, GroupConv2D layers with kernel sizes of 3×3, 5×5, 7×7, and 11×11 split the feature channels into multiple groups, each capturing different localized features. Due to the complexity of channel conditions, predicting future channels is challenging because the locations of useful features vary significantly over time. By utilizing a multi-branch Inception architecture, the cascaded modules extract both local and global features from the codewords. In the final block, outputs from convolution layers with varying kernel sizes are fused through addition, integrating multiple spatio-temporal CSI features at different scales.

Note that the joint CSI feedback and prediction model is trained in an end-to-end manner. Its parameters are updated using an adaptive moment estimation (ADAM) optimizer. The networks are trained to minimize the difference between the predicted and the ground truth CSIs. Consequently, the training loss function is defined as the mean square error (MSE) expressed as follows.

$$L(\theta_{\text{en}},\theta_{\text{pre}},\theta_{\text{de}}) = \frac{1}{T}\sum_{i=1}^{T}\sum_{j=1}^{L}\left\| \boldsymbol{H}_i^{(t+j)} - \bar{\boldsymbol{H}}_i^{(t+j)} \right\|^2 \quad (7),$$

where T is the number of samples in the training set, and the subscript of $\boldsymbol{H}$ denotes the i-th sample in the training set. $\boldsymbol{H}_i^{(t+j)}$ and $\bar{\boldsymbol{H}}_i^{(t+j)}$ denote actual and predicted CSI at the $(t+j)$-th slot, respectively.

3.2 SNR Adaptive Module

In high-speed mobile scenarios, the uplink feedback channel undergoes rapid variations, requiring the end-to-end feedback system to adapt automatically to changing channel conditions. To address this, we introduce an SNR adaptive module (SAM), depicted in Fig. 3. The SAM is designed based on the mechanism of channel-wise soft attention[19], which identifies channel relationships and generates distinct scaling parameters for different channel states, thereby enhancing or attenuating their influence on subsequent layers[20]. By dynamically adjusting resource allocation strategies based on these varying channel states, the system implicitly modulates the source coding rates in both the encoder and decoder, ultimately achieving higher-quality transmission and CSI reconstruction.

As illustrated in Fig. 3, the SAM consists of three components: 1) SNR semantic extraction, 2) semantic embedding, and 3) feature calibration. The channel feature s is first processed by the fully connected (FC) layer and then fed into the SAM for modulation.

1) SNR semantic extraction. The uplink channel information SNR is first input into the three FC layers to generate the semantic information of the channel state. The first and second FC layers are followed by the Rectified Linear Unit (ReLU) and the last FC layer is followed by a sigmoid to restrict the output range to the interval (0, 1)[21]. It transforms SNR into an M-dimensional vector $\boldsymbol{v}_{\text{SNR}}$.

2) Semantic embedding. The input channel features and the extracted SNR semantic information $\boldsymbol{v}_{\text{SNR}}$ are fused and embedded by the element-wise product. The output will pass through the next FC layer and continue to be multiplied by $\boldsymbol{v}_{\text{SNR}}$. Following three rounds of semantic embedding, it will be restored to the same channel dimension as s via the last FC layer, and then pass through a sigmoid function to obtain the modulation scale factor.

3) Feature calibration. The resulting modulation scale factor is subsequently multiplied by the original channel characteristics to obtain the CSI feature map s'.

The SNR adaptive module integrates the SNR directly into the token processing pipeline to compute channel-wise attention, enhancing the adaptability of the network in scenarios with varying signal conditions[22]. Algorithm 1 summarizes the operation process of the proposed SAM.

Algorithm 1. Operation process of SAM

Input: The channel feature s and the uplink channel SNR
Output: The calibrated channel feature s'
1. Upgrade the channel features to M dimensions and get s_M
2. Extract the SNR semantic vector: $\boldsymbol{v}_{\text{SNR}} = \text{Sigmoid}(W_3\text{ReLU}(W_2\text{ReLU}(W_1\text{SNR} + b_1) + b_2) + b_3)$
3. Combine features and the SNR semantic vector: $\text{output}_0 = s_M \cdot \boldsymbol{v}_{\text{SNR}}$
4. **For** $i = 1 : 1 : 3$ **do**
5. Embed SNR semantic information in features: $\text{output}_i = (W_{Mi}\text{output}_{i-1} + b_{Mi}) \cdot \boldsymbol{v}_{\text{SNR}}$
6. **end for**
7. Calculate the modulation scale factor: $\mu = \text{Sigmoid}(W_C\text{output}_3 + b_C)$

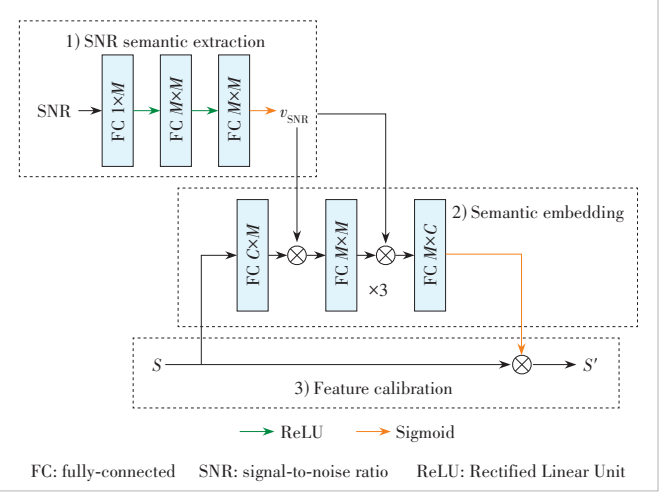


Figure 3. Architecture of SAM

8. Obtain the calibrated channel feature: $s' = s \cdot \mu$
9. **return** s'

4 Experimental Results

In this section, we present the numerical results to investigate the performance of the proposed STPNet design for joint CSI feedback and prediction.

4.1 Experiment Settings

The simulation results are based on the clustered delay line (CDL)-C channel model and the 3GPP urban macro (UMa) channel model[23], respectively. The BS employs a uniform rectangular panel array of dual-polarized antennas arranged in an 8×2 configuration. The user speed is set to 30 km/h. There are N_c = 32 subcarriers with 10 MHz bandwidth. The communication frequency f is set as 2 GHz. The lengths of historical and predicted CSIs are both set to 5. Table 1 summarizes the simulation parameters. The training and testing datasets contain 10 000 and 2 000 samples, respectively. To enhance model generalization, the prediction model is trained using uplink channels with SNR values ranging from 1 dB to 20 dB. We update the parameters with a constant learning rate of 1×10^{-3}. The batch size and the training epoch are set as 16 and 100, respectively. To evaluate model effectiveness, we quantify the accuracy of CSI prediction by using normalized mean square error (NMSE) as a quantitative metric. The NMSE is defined as:

$$\mathrm{NMSE} = \mathbb{E}\left(\frac{\left\| \boldsymbol{H} - \bar{\boldsymbol{H}} \right\|^2}{\left\| \boldsymbol{H} \right\|^2}\right) \tag{8},$$

where $\boldsymbol{H} \in \mathbb{C}^{L \times N_c \times N_t}$ denotes the ideal channel for the next L slots, and $\bar{\boldsymbol{H}} \in \mathbb{C}^{L \times N_c \times N_t}$ denotes the predicted channel.

Fig. 4a shows a sample from a single BS antenna selected for simulation from the CDL-C scenario CSI dataset. The duration of this particular sample is 10 ms. The time-varying nature of the wireless channel is captured by its autocorrelation function (ACF), as illustrated in Fig. 4b. This second-order statistic is typically influenced by factors such as the propagation geometry, the mobile's velocity, and the characteristics of the antennas[24–25]. In this paper, the DL-based approach is adopted to learn and capture the spatio-temporal correlation of CSI.

Table 1. Simulation parameters

Parameter	Value
Channel type	3GPP CDL-C and UMa[23]
Carrier frequency	2 GHz
Bandwidth	10 MHz
N_t	32
N_r	1
Number of subcarriers	32
Feedback interval	1 ms
UE speed	30 km/h

CDL: clustered delay line UE: user equipment UMa: urban macro

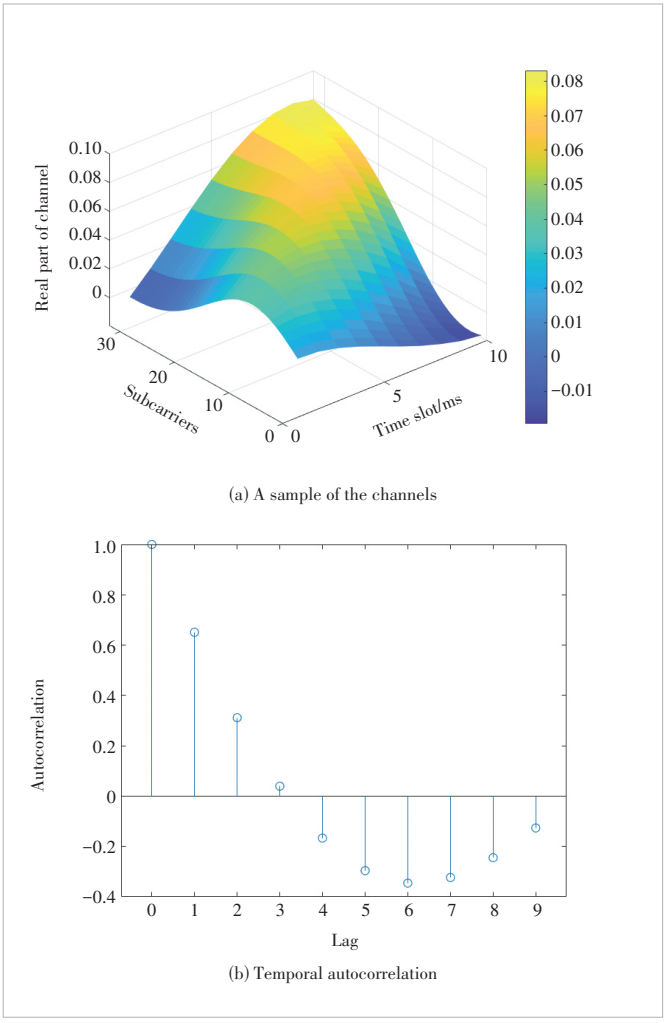
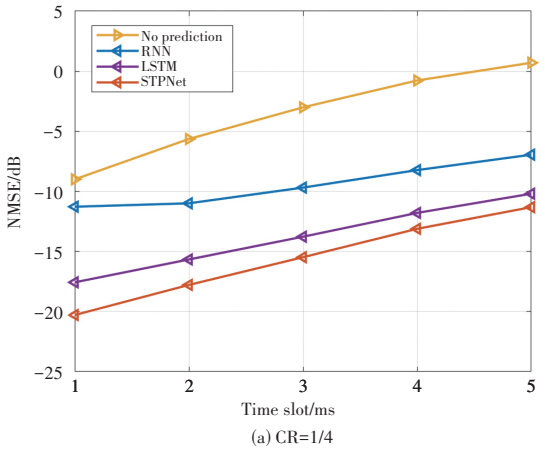


Figure 4. A sample from the CDL-C channel model CSI dataset and the temporal autocorrelation

4.2 Performance Comparison

We primarily compare our CSI prediction module with some existing ones, such as the RNN-based method[13] and the LSTM-based method[14]. To ensure a fair comparison, all baseline prediction methods are jointly trained with the CSI feedback network. The CSI feedback process is implemented using the SwinCFNet architecture with an SNR-adaptive module. Fig. 5 demonstrates the NMSE performance of the proposed and baseline methods at CR=1/4, 1/8 in the CDL-C channel model. The test SNR is set to 20 dB. The performance of non-prediction schemes represents the gaps between the reconstructed nearest historical CSI and the future CSI, which further underscores the importance of channel prediction in the feedback process.



LSTM: long short-term memory
NMSE: normalized mean square error
RNN: recurrent neural network
STPNet: spatio-temporal predictive network

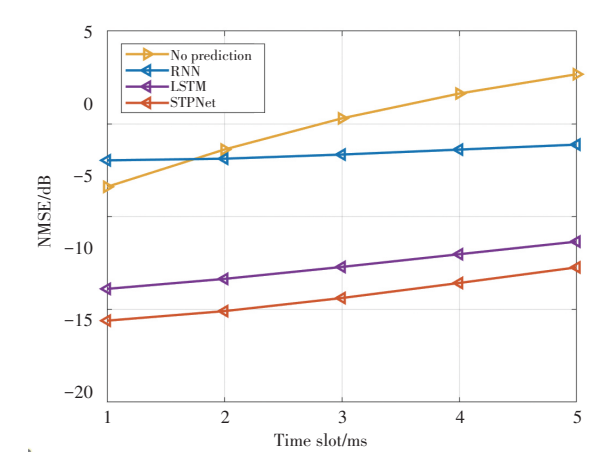
Figure 5. NMSE performance in the CDL-C channel model with CR=1/4 and 1/8

From both Figs. 5a and 5b, it is seen that the NMSE performance of all evaluated algorithms decreases over time. As illustrated in Fig. 5, the proposed Inception-based STPNet achieves the highest performance in the CDL-C channel. For example, when CR is equal to 1/4, STPNet attains NMSE gains of 6.79 dB and 2.12 dB over the RNN-based and LSTM-based methods, respectively, when predicting the channel at the second future slot. Furthermore, compared with the non-prediction scenario, STPNet improves the accuracy of the fifth time slot by more than 12 dB at CR = 1/8. Under these settings, STPNet also achieves an additional 1.43 dB NMSE improvement over the best results of other competing methods.

The improvements of the proposed channel prediction scheme in Fig. 5 come from two aspects. First, the traditional RNN-based prediction methods operate recursively, using the current time slot as input to predict the next. While effective for short-term forecasting, this approach often leads to substantial performance degradation when extrapolating over extended future intervals. In contrast, our proposed scheme predicts all future channels simultaneously, thereby breaking the recursive loop and preventing error accumulation. Second, rather than treating CSI as a time series, our method represents it as a spatial map, capturing the spatio-temporal correlations embedded in the data. By leveraging a multi-branch architecture, the Inception-based CSI prediction module effectively extracts both local and global features from stacks of temporal dynamics.

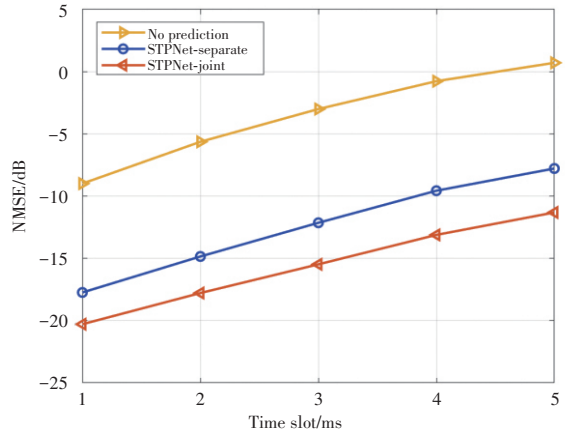
In Fig. 6, we compare the NMSE performance of STPNet and other prediction networks with CR=1/4 in the UMa channel model generated on QuaDRiGa[26]. The test SNR is set to 20 dB. Since the 3GPP UMa model randomly samples channel parameters, the resulting channels exhibit greater randomness and reduced predictability compared with the CDL-C model. Nevertheless, as shown in Fig. 6, STPNet maintains the state-of-the-art NMSE performance. Notably, for the prediction of the channel at the first future slot, the RNN-based method proves less accurate than the non-prediction approach due to the gradient vanishing problem. Compared with the non-prediction scheme and the LSTM-based method, STPNet achieves improvements in NMSE of 80.99% and 32.56%, respectively.

Furthermore, we investigate the performance of joint CSI feedback and prediction compared with separate CSI feedback and prediction. In the STPNet-separate configuration, CSI feedback and channel prediction networks are trained independently and then evaluated in series. As illustrated in Fig. 7a, the joint architecture, STPNet-joint, achieves at least a 2 dB improvement in NMSE over the STPNet-separate configuration, demonstrating the effectiveness of joint training. Fig. 7b shows the achievable sum-rate performance of different methods. The upper bound is attained by the scheme with perfect channel information available. We can also observe that STPNet-joint could approximate the near-optimal sum-rate performance attained with perfect channel information. For instance, when pre-

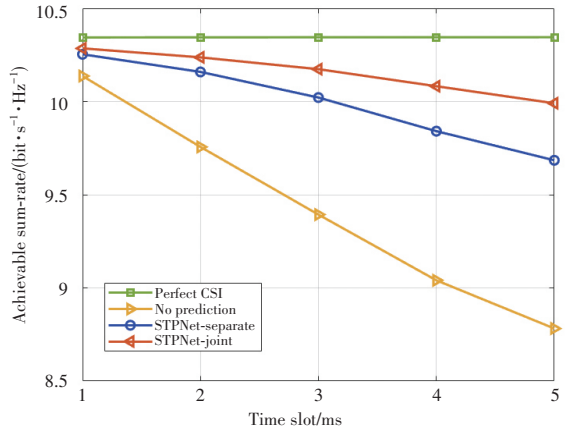


LSTM: long short-term memory
NMSE: normalized mean square error
RNN: recurrent neural network
STPNet: spatio-temporal predictive network

Figure 6. NMSE performance in the UMa channel model with CR=1/4



(a) NMSE performance



(b) Achievable sum-rate performance

CSI: channel state information
LSTM: long short-term memory
NMSE: normalized mean square error
RNN: recurrent neural network
STPNet: spatio-temporal predictive network

Figure 7. NMSE and achievable sum-rate performance of different architectures in the CDL-C channel model with CR=1/4

dicting the channel for the fifth future slot, STPNet-joint achieves approximately 96.57% of the sum-rate performance of the upper bound. By integrating CSI feedback and prediction, the system avoids error propagation between these two cascaded subsystems, thereby enhancing overall accuracy.

5 Conclusions

This paper presents STPNet, an efficient spatio-temporal predictive network based on a joint feedback and prediction framework. STPNet is designed to address the challenges of excessive feedback overhead and dynamic channel conditions in massive MIMO systems. The CSI prediction module is stacked with a series of Inception modules used for capturing both the local and global spatio-temporal features. By leveraging spatio-temporal features and SNR-aware modulation, STPNet achieves outstanding performance in CSI prediction accuracy and robustness, significantly outperforming traditional methods. Simulation results validate the effectiveness of the proposed framework across diverse channel scenarios, demonstrating its potential to enhance future wireless communication systems. Future work will explore extending the model to more complex and dynamic environments, further improving its adaptability and efficiency.

References

[1] CHEN W S, LIN X Q, LEE J, et al. 5G-advanced toward 6G: past, present, and future [J]. IEEE journal on selected areas in communications, 2023, 41 (6): 1592 – 1619. DOI: 10.1109/JSAC.2023.3274037
[2] CHEN W, LIU Y W, JAFARKHANI H, et al. Signal processing and learning for next generation multiple access in 6G [J]. IEEE journal of selected topics in signal processing, 2024, 18(7): 1146 – 1177. DOI: 10.1109/JSTSP.2024.3511403
[3] WEN C K, SHIH W T, JIN S. Deep learning for massive MIMO CSI feedback [J]. IEEE wireless communications letters, 2018, 7(5): 748 – 751. DOI: 10.1109/LWC.2018.2818160
[4] GAO Y CHEN J J, LI D P. Intelligence driven wireless networks in B5G and 6G era: a survey [J]. ZTE communications, 2024, 22(3): 99 – 105. DOI: 10.12142/ZTECOM.202403012
[5] YANG B LIANG X, LIU S N, et al. Intelligent 6G wireless network with multi-dimensional information perception [J]. ZTE communications, 2023, 21(2): 3 – 10. DOI: 10.12142/ZTECOM.202302002
[6] GUO Y R, CHEN W, SUN F F, et al. Deep learning for CSI feedback: one-sided model and joint multi-module learning perspectives [EB/OL]. (2024-05-09)[2024-12-12]. http://export.arxiv.org/abs/2405.05522
[7] CHENG J M, CHEN W, XU J L, et al. Swin Transformer-based CSI feedback for massive MIMO [C]//The 23rd International Conference on Communication Technology (ICCT). IEEE, 2023: 809 – 814. DOI: 10.1109/ICCT59356.2023.10419637
[8] YI X P, YANG S, GESBERT D, et al. The degrees of freedom region of temporally correlated MIMO networks with delayed CSIT [J]. IEEE transactions on information theory, 2014, 60(1): 494 – 514. DOI: 10.1109/TIT.2013.2284500
[9] MIHAI S, YAQOOB M, HUNG D V, et al. Digital twins: a survey on enabling technologies, challenges, trends and future prospects [J]. IEEE communications surveys & tutorials, 2022, 24(4): 2255 – 2291. DOI: 10.1109/COMST.2022.3208773
[10] TAN J SHA X B, DAI B, et al. Analysis of industrial Internet of Things and digital twins [J]. ZTE communications, 2021, 19(2): 53 – 60. DOI: 10.12142/ZTECOM.202102007
[11] YIN H F, WANG H Q, LIU Y Z, et al. Addressing the curse of mobility in massive MIMO with prony-based angular-delay domain channel predictions [J]. IEEE journal on selected areas in communications, 2020, 38 (12): 2903 – 2917. DOI: 10.1109/JSAC.2020.3005473
[12] BADDOUR K E, BEAULIEU N C. Autoregressive modeling for fading channel simulation [J]. IEEE transactions on wireless communications, 2005, 4(4): 1650 – 1662. DOI: 10.1109/TWC.2005.850327
[13] JIANG W, SCHOTTEN H D. Neural network-based fading channel prediction: A comprehensive overview [J]. IEEE access, 2019, 7: 118112 – 118124. DOI: 10.1109/ACCESS.2019.2937588
[14] JIANG W, SCHOTTEN H D. Deep learning for fading channel prediction [J]. IEEE open journal of the communications society, 2020, 1: 320 – 332. DOI: 10.1109/OJCOMS.2020.2982513
[15] JIANG H, CUI M Y, NG D W K, et al. Accurate channel prediction based on transformer: making mobility negligible [J]. IEEE journal on selected areas in communications, 2022, 40(9): 2717 – 2732. DOI: 10.1109/JSAC.2022.3191334
[16] REN Z Z, ZHANG X D, WANG J T. Joint CSI feedback and prediction with deep learning in high-speed scenarios [C]//Proceedings of IEEE/CIC

International Conference on Communications in China (ICCC). IEEE, 2024: 1910 – 1915. DOI: 10.1109/ICCC62479.2024.10681972
[17] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions [C]// Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015: 1 – 9. DOI: 10.1109/CVPR.2015.7298594
[18] GAO Z Y, TAN C, WU L R, et al. SimVP: Simpler yet better video prediction [C]//Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 3160 – 3170. DOI: 10.1109/CVPR52688.2022.00317
[19] XU J L, AI B, CHEN W, et al. Wireless image transmission using deep source channel coding with attention modules [J]. IEEE transactions on circuits and systems for video technology, 2022, 32(4): 2315 – 2328. DOI: 10.1109/TCSVT.2021.3082521
[20] XU J L, AI B, WANG N, et al. Deep joint source-channel coding for CSI feedback: an end-to-end approach [J]. IEEE journal on selected areas in communications, 2023, 41(1): 260 – 273. DOI: 10.1109/JSAC.2022.3221963
[21] YANG K, WANG S X, DAI J C, et al. WITT: a wireless image transmission transformer for semantic communications [C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1 – 5. DOI: 10.1109/ICASSP49357.2023.10094735
[22] DENG L T ZHAO Y R. Deep learning-based semantic feature extraction: a literature review and future directions [J]. ZTE communications, 2023, 21(2): 11 – 17. DOI: 10.12142/ZTECOM.202302003
[23] 3GPP. Study on channel model for frequencies from 0.5 to 100 GHz: TR 38.901 V18.0.0 [S]. 2024
[24] WU C, YI X P, ZHU Y M, et al. Channel prediction in high-mobility massive MIMO: from spatio-temporal autoregression to deep learning [J]. IEEE journal on selected areas in communications, 2021, 39(7): 1915 – 1930. DOI: 10.1109/JSAC.2021.3078503
[25] YUAN J D, NGO H Q, MATTHAIOU M. Machine learning-based channel prediction in massive MIMO with channel aging [J]. IEEE transactions on wireless communications, 2020, 19(5): 2960 – 2973. DOI: 10.1109/TWC.2020.2969627
[26] JAECKEL S, RASCHKOWSKI L, BÖRNER K, et al. Quasi deterministic radio channel generator, user manual and documentation [R]. Berlin, Germany: QuaDRiGa, 2021

Biographies

CHENG Jiaming received his BE degree from Beijing Jiaotong University, China in 2024, where he is currently pursuing his PhD degree. His current research interests include massive MIMO and intelligent communications.

CHEN Wei (weich@bjtu.edu.cn) received his BE and ME degrees from the Beijing University of Posts and Telecommunications, China in 2006 and 2009, respectively, and PhD degree in computer science from the University of Cambridge, UK in 2013. Later, he was a research associate with the Computer Laboratory, University of Cambridge, from 2013 to 2016. He is currently a professor with Beijing Jiaotong University, China. He has published over 130 articles and won several international awards. His current research interests include intelligent wireless communication systems and multimedia processing.

LI Lun received his MS degree in electronics and communication engineering from Harbin Institute of Technology, China in 2018. He joined ZTE Corporation in 2018, where he is currently a technical pre-research engineer. His research interests include artificial intelligence/machine learning for wireless communications.

AI Bo received his MS and PhD degrees from Xidian University, China in 2002 and 2004, respectively. He is currently a full professor with Beijing Jiaotong University, China. He has authored/coauthored eight books and published over 300 academic research articles. His research interests include the research and applications of channel measurement and channel modeling, and dedicated mobile communications for rail traffic systems. He has received many awards, such as the Distinguished Youth Foundation and the Excellent Youth Foundation from the National Natural Science Foundation of China, the Qiushi Outstanding Youth Award by Hong Kong Qiushi Foundation, the New Century Talents by the Chinese Ministry of Education, the Zhan Tianyou Railway Science and Technology Award by the Chinese Ministry of Railways, and the Science and Technology New Star Award by the Beijing Municipal Science and Technology Commission.

RIS Enabled Simultaneous Transmission and Key Generation with PPO: Exploring Security Boundary of RIS Phase Shift

FAN Kaiqing[1], YAO Yuze[1], GAO Ning[1], LI Xiao[2], JIN Shi[2]

(1. School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China;
2. National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China)

DOI: 10.12142/ZTECOM.202501003

https://kns.cnki.net/kcms/detail/34.1294.TN.20250226.1147.002.html, published online February 26, 2025

Manuscript received: 2025-01-25

Abstract: Due to the broadcast nature of wireless channels and the development of quantum computers, the confidentiality of wireless communication is seriously threatened. In this paper, we propose an integrated communications and security (ICAS) design to enhance communication security using reconfigurable intelligent surfaces (RIS), in which the physical layer key generation (PLKG) rate and the data transmission rate are jointly considered. Specifically, to deal with the threat of eavesdropping attackers, we focus on studying the simultaneous transmission and key generation (STAG) by configuring the RIS phase shift. Firstly, we derive the key generation rate of the RIS assisted PLKG and formulate the optimization problem. Then, in light of the dynamic wireless environments, the optimization problem is modeled as a finite Markov decision process. We put forward a policy gradient-based proximal policy optimization (PPO) algorithm to optimize the continuous phase shift of the RIS, which improves the convergence stability and explores the security boundary of the RIS phase shift for STAG. The simulation results demonstrate that the proposed algorithm outperforms the benchmark method in convergence stability and system performance. By reasonably allocating the weight factors for the data transmission rate and the key generation rate, "one-time pad" communication can be achieved. The proposed method has about 90% performance improvement for "one-time pad" communication compared with the benchmark methods.

Keywords: reconfigurable intelligent surfaces; physical layer key generation; integrated communications and security; one-time pad; deep reinforcement learning

Citation (Format 1): FAN K Q, YAO Y Z, GAO N, et al. RIS enabled simultaneous transmission and key generation with PPO: exploring security boundary of RIS phase shift [J]. ZTE Communications, 2025, 23(1): 11 - 17. DOI: 10.12142/ZTECOM.202501003
Citation (Format 2): K. Q. Fan, Y. Z. Yao, N. Gao, et al., "RIS enabled simultaneous transmission and key generation with PPO: exploring security boundary of RIS phase shift," *ZTE Communications*, vol. 23, no. 1, pp. 11 - 17, Mar. 2025. doi: 10.12142/ZTECOM.202501003.

1 Introduction

With the advancement of the 6G wireless communication, we are gradually moving towards the era of comprehensive Internet of Things (IoT). This provides more solid technical support for applications like smart interaction, industrial control, and remote healthcare, which requires extremely low latency while ensuring high security[1]. However, the widespread access to diversified intelligent mobile terminals and the demand for Gbit/s-level ultra-high throughput highlight the crucial importance of data security, especially in the broadcast wireless channels. Traditional key encryption methods may not be able to meet such stringent security requirements[2]. Meanwhile, it is necessary to meet the requirements for rapid key generation to reduce communication latency and ensure key security to prevent from cracking by quantum computers. This necessitates an in-depth exploration of the key distribution mechanism to discover the optimal trade-off between latency and security. Hence, in the 6G era, constructing a secure and efficient confidential communication system is urgently demanded.

In recent years, physical layer key generation (PLKG) has garnered increasing attention in academics and industry. PLKG is based on the physical layer characteristics of wireless environments, including the wireless channels that inherently possess randomness and reciprocal features. PLKG leverages these characteristics to establish a key generation mechanism, thus circumventing the challenges of traditional key distribution and update approaches. Typically, PLKG encompasses four steps[3]. First comes channel sounding, where the transceiver sends a pilot sequence to detect the channel and obtain reciprocal characteristics. Next is the quantization step, where the channel reciprocity features are transformed

This work was supported in part by the National Science Foundation of China (NSFC) under Grant No. 62371131, in part by the National Key R&D Program of China under Grant No. 2024YFE0200700, and in part by the program of Zhishan Young Scholar of Southeast University under Grant No. 2242024RCB0030.

into a binary bit sequence, and this bit sequence is then generated as the raw key. Due to issues like quantization accuracy, noise, and incomplete synchronization, the original bit sequence might not match properly. The third step is information reconciliation, where the error correcting codes are employed for correction purposes. Finally, privacy amplification is utilized, which aims to eliminate the potential risks of information leakage within the original bit sequence and generate symmetric keys to safeguard data security. MAURER[4] first explores the problem of generating shared keys through public discussions when both parties are aware of the relevant random variables but do not have an initial shared key. PREMNATH et al. evaluate the effectiveness of extracting keys from changes in wireless signal strength through actual measurements in Ref. [5]. They find that there are some problems with key generation in poor scattering environments, e.g., the entropy of the key is relatively low and the attacker can easily crack the key. An adaptive key generation scheme has been proposed to address these issues. In Ref. [6], LI et al. focus on using principal component analysis (PCA) preprocessing to generate highly consistent uncorrelated keys. However, due to the low-key generation rate in static wireless environments like an indoor office, it seriously affects the key generation rate.

At present, some related studies begin to focus on reconfigurable intelligent surfaces (RIS) assisted PLKG to improve the key generation rate[7]. For example, Ref. [8] proposes a RIS assisted multi-carrier physical layer key generation framework to address the issue of insufficient randomness in wireless channels in static environments. Ref. [9] proposes the "SemKey" scheme, which utilizes the semantic drift phenomenon in semantic communication systems combined with RIS assistance to improve the key generation rate. The advantages and feasibility of this scheme have been experimentally verified. Ref. [10] proposes a RIS configuration method that utilizes channel state information (CSI) to control the activation of specific RIS units in the presence of eavesdroppers, thereby increasing the key capacity. However, the robust security of 6G enabled by the the RIS assisted PLKG, i.e, achieving "one-time pad" communications, still needs further study.

In 6G, the density of IoT devices per square kilometer can reach over 10 million. In such massive connection scenarios, communication security is extremely vulnerable. Integrated communications and security (ICAS) provides a potential solution to strong security, which shares communication resources and hardware resources and conducts an integrated design of communication functions and security functions. Specifically, the inherent by-products of communication are utilized to enhance the security abilities; at the same time, the improvement of security capabilities further ensures communication security, thereby enabling communication and security to mutually benefit and be internally generated with each other[2, 11]. Since the ICAS design focuses on real-time extreme security communication, artificial intelligence (AI) is an important endogenous power, especially deep learning (DL) and reinforcement learning (RL). In dynamic wireless environments, GAO et al. use deep Q-network (DQN) to optimize the RIS phase shift and for the first time demonstrate that the simultaneous transmission and key generation (STAG) can achieve "one-time pad" communication[11]. However, the existing DQN-based STAG method has some drawbacks, including the dimension explosion problem when the action space is large, and poor performance when there are many RIS units or high phase shift resolution. On the other hand, with the improvement of the RIS hardware manufacturing process, the high performance RIS with 3 bits or higher resolution, e.g., 360 degree RIS, has gradually emerged[12]. Motivated by these considerations, we propose a proximal policy optimization (PPO) based STAG method to study the security boundary of RIS phase shifts for STAG. The main contributions are summarized as follows.

- To improve the convergence stability of the deep reinforcement learning (DRL)-based STKG, a PPO-based STAG method is proposed. In particular, the RIS-assisted key generation rate is derived and the triple of the DRL, i.e., action, state, and reward, with respect to the STAG, is constructed.
- The continuous phase shift of RIS is optimized to explore the security boundary of RIS phase shifts. The upper bound of the RIS phase shifting capability for STAG is evaluated via the simulation. The continuous RIS phase shift yields over 5% higher reward than the 1-bit discrete RIS phase shift when the proposed algorithm converges.
- The simulation result shows that the "one-time pad" communication can be achieved by assigning suitable weight factors to STAG. Compared with the DQN-based method, the proposed PPO-based STAG method can obtain 90% performance improvement in "one-time pad" communication.

2 System Model

In Fig. 1, we consider a static RIS-assisted key generation scenario, which consists of four components: the legitimate transmitter and the receiver, namely Alice and Bob, the RIS,

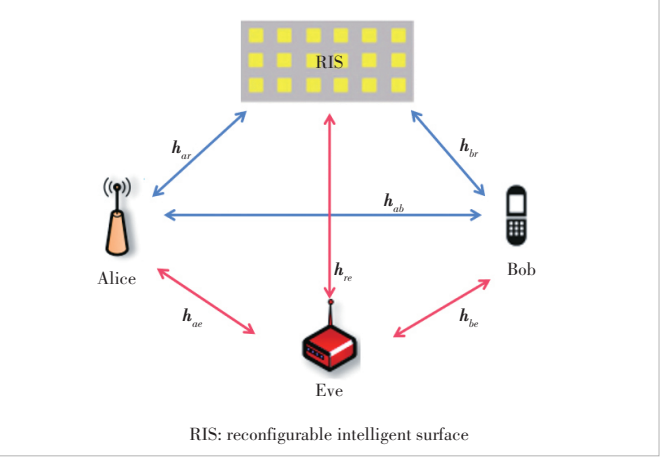


Figure 1. System model schematic diagram

and the malicious eavesdropper Eve. To simplify the analysis, we assume that Eve is in the middle of the legitimate users. Each participant is location-fixed and equipped with one antenna, and RIS has N reflection units.

2.1 Channel Model

The transmitter and the receiver intend to simultaneously generate keys and transmitted data, and the eavesdropper passively eavesdrops on the channel information. The signal received by Alice can be represented as:

$$r_a = \underbrace{\left(h_{br}^T \boldsymbol{\Phi} h_{ra} + h_{ba}\right)}_{h_{bra}} s_b + n_a \tag{1},$$

where $h_{br} \in C^{N\times 1}$ is the channel from Bob to RIS, $h_{ra} \in C^{1\times N}$ is the channel from RIS to Alice, $h_{ba} \in C$ is the direct channel from Bob to Alice, h_{bra} is the equivalent channel, s_b is the transmission signal of Bob, $\boldsymbol{\Phi} = \mathrm{diag}[\alpha_1 e^{j\theta_1}, \alpha_2 e^{j\theta_2}, \cdots, \alpha_N e^{j\theta_N}]$ is the phase-shift matrix of RIS with $\phi_{n,n} = \alpha_n e^{j\theta_n}$, $\alpha_n = 1, \theta_n \in [0, 2\pi)$, $n = 1,2,\cdots,N$, and n_a is the channel noise following complex Gaussian distribution with zero mean and σ^2 variance.

Similarly, we can obtain the received signals of Bob and Eve, respectively, which is given by Eqs. (2) and (3). Therein, $h_{re} \in C^{N\times 1}$ is the channel from RIS to Eve, $h_{ae} \in C^{N\times 1}$ is the channel from Alice to Eve, and n_a and n_e are the channel noise.

$$r_b = \underbrace{\left(h_{ar}^T \boldsymbol{\Phi} h_{rb} + h_{ab}\right)}_{h_{arb}} s_a + n_b \tag{2},$$

$$r_e = \underbrace{\left(h_{ar}^T \boldsymbol{\Phi} h_{re} + h_{ae}\right)}_{h_{are}} s_a + n_e \tag{3}.$$

2.2 Channel Estimation

During the coherent time, Alice exchanges the pilot signal with Bob for channel estimation. Let Alice be the communication initiator, and Bob estimates CSI through least squares.*

$$\hat{h}_{arb} = h_{ar}^T \boldsymbol{\Phi} h_{rb} + h_{ab} + n_0 s_a^{p*} \tag{4},$$

where s_a^p is the pilot signal from Alice to Bob and the pilot signal satisfies $s_a^p s_a^{p*} = 1$; the channel estimation error is $n_0 s_a^{p*}$. Next, symmetric keys are generated through quantization, information reconciliation and privacy amplification[13]. Since these steps are not the key point of this paper, they are not elaborated on any further.

2.3 Key Generation Rate

The mutual information between the channel observations of the legitimate parties is an important factor in determining the key generation rate. Due to quantization error in bit representation, we consider the mutual information as the upper bound of the key generation rate, which is the mutual information of CSI under Eve's observation. With the eavesdropper Eve, the key generation rate can be formulated as[14]:

$$\begin{aligned}\mathcal{R}_{\mathrm{key}} &= \frac{1}{T} I(\hat{h}_{arb}; \hat{h}_{bra} | \hat{h}_{are}) = \\ &\frac{1}{T}\left[H(\hat{h}_{arb}|\hat{h}_{are}) - H(\hat{h}_{arb}|\hat{h}_{bra}, \hat{h}_{are})\right] = \\ &\frac{1}{T}\log_2 \frac{\det(R_{ae})\det(R_{\mathrm{be}})}{\det(R_e)\det(R_{abe})}\end{aligned} \tag{5},$$

where $\det(R)$ is the matrix determinant, while $\boldsymbol{R}_{ae}, \boldsymbol{R}_{be}, \boldsymbol{R}_e$, and $\boldsymbol{R}_{abe}$ are the covariance matrices. T represents the observation time. Specifically, the covariance matrix is as follows:

$$\boldsymbol{R}_{A_1,\cdots,A_n} = E\begin{bmatrix} a_1 a_1^* & \cdots & a_1 a_n^* \\ \vdots & \ddots & \vdots \\ a_n a_1^* & \cdots & a_n a_n^* \end{bmatrix} \tag{6}.$$

The key rate is expressed in Eq. (7), where $\|\cdot\|$ is the Euclidian norm operator. E represents mathematical expectation. For convenience, we simplify the variance of the noise to 1. Thus, we can obtain the key generation rate, which is expressed in Eq. (8).

$$\mathcal{R}_{\mathrm{key}} = \log_2\left(\frac{((R_a+\sigma^2)(R_e+\sigma^2) - \|R_{ae}\|^2)^2}{(R_e+\sigma^2)((2R_a\sigma^2+\sigma^4)(R_e+\sigma^2) - 2\sigma^2\|R_{ae}\|^2)}\right) \tag{7},$$

$$\mathcal{R}_{\mathrm{key}} = \log_2\left(\frac{((R_a+1)(R_e+1) - \|R_{ae}\|^2)^2}{(R_e+1)((2R_a+1)(R_e+\sigma^2) - 2\|R_{ae}\|^2)}\right) \tag{8}.$$

According to Shannon's formula, the maximum channel capacity is the theoretical maximum transmission rate, which can be obtained by calculating the signal-to-noise ratio (SNR). Thus, we can obtain the ergodic data transmission rate at Alice as:

$$\mathcal{R}_{\mathrm{data}} = B\log_2\left(1 + E\|h_{br}^T \boldsymbol{\Phi} h_{ra} + h_{ba}\|^2\right) \tag{9},$$

where B is the signal bandwidth.

3 Problem Formulation and Proposed Solution

3.1 Problem Formulation

We consider jointly optimizing the key generation rate and data transmission rate, that is, to ensure the data transmission

* The channel estimation considered in this paper has no error, and the analysis is based on perfect channel state information. The research based on imperfect channel state information will be carried out in the future.

rate reaches a high level while maximizing the key generation rate to enhance the confidentiality of wireless communication. For the trade-off between the key generation rate and the data transmission rate, we make decisions based on the specific application scenarios, such as in real-time communication prioritized applications about the voice calls and the video conferences, which increases the weight of the data transmission rate and appropriately reduces the key generation rate. For financial transaction scenarios and military communication scenarios that focus on high security and confidentiality, we increase the weight of the key generation rate accordingly. In short, we can first evaluate the security and the quality of service (QoS) requirements of the scenario and allocate corresponding weight reasonably to the specific scenario. Therefore, we can formulate the optimization problem as

$$\begin{aligned}&\mathbb{P}: \quad \max\ w_d\mathcal{R}_{\text{data}} + w_k\mathcal{R}_{\text{key}}\\&\text{s.t.}\ \ 0 \leqslant \theta_n < 2\pi, \forall n \in \{1,\cdots,N\}\\&|\phi_{n,n}| = 1\end{aligned} \tag{10}$$

where $w_d \in [0,1]$, $w_k = 1 - w_d \in [0,1]$ is the weight that balances the priority level of the key generation rate and data transmission, n represents the number of reflection units of RIS, θ_n represents the phase shift unit of RIS, and $|\phi_{n,n}|$ is the phase-shift unit of RIS with a constant modulus constraint.

Due to the non-convex nature of the optimization problem, it is hard for the traditional convex optimization to obtain the optimal solution in real-time. Considering the dynamic wireless environments, we construct the time series of the dynamic channel as a Markov decision process. This indicates that DRL is a potent instrument for resolving the Markov decision process. PPO is a model-free reinforcement learning algorithm, which belongs to the family of strategy gradient algorithms. It is mainly used to optimize the strategy network so that the agents can take optimal actions in the environment to maximize the cumulative rewards. Due to the increasing demand for efficient and stable algorithms, PPO has emerged where the action space is continuous. It not only performs well in the large dimensional action space but also has the advantages of high training efficiency and easy convergence. Therefore, we use the PPO algorithm to jointly optimize the transmission rate and the key generation rate with the continuous RIS phase shift[15].

3.2 Sample Collection

Firstly, we use the current strategy network to interact with the environment and collect a series of state-action-reward samples $\{(s_i,a_i,r_i)\}$[16]. These samples form an experience replay buffer. Then, the advantage function and target value are calculated based on the collected samples, and the state value function is estimated to calculate the advantage function $A(s,a)$. Here, we use Monte Carlo estimation to calculate the value function, where the advantage function can be calculated by subtracting the state value function from the cumulative reward of the trajectory[17]. For time difference learning, it can be denoted as:

$$A(s,a) = r + \gamma V(s') - V(s) \tag{11}$$

where r is the instant reward, γ is the discount factor, and s' is the next state.

3.3 Strategy Network Update

In this step, the gradient descent algorithm is employed to minimize the loss function and optimize the policy network. The loss function $L^{\text{CLIP}}(\theta)$ is calculated to obtain the gradient of the policy network parameter. Then, the gradient descent is used to update via the formula $\theta = \theta - \beta\nabla_\theta L^{\text{CLIP}}(\theta)$, where β represents the learning rate[18]. The specific settings of the state space, the action space, and the reward function in the Markov decision process are as follows.

State: The state space is defined as the CSI of the communication environment observed by Alice. Therefore, at time step i, the state is denoted as:

$$s^i = \left\{h^i_{bar,\Phi^{i-1}}, h^i_{bae,\Phi^{i-1}}\right\} \tag{12}$$

The state information is the basis for intelligent agents to make decisions.

Action: Since we train the network by continuously adjusting the phase shift of RIS, the action space at time step i can be represented as:

$$a^i = \{\boldsymbol{\Phi}^i\} \tag{13}$$

where $\boldsymbol{\Phi} = \text{diag}[\alpha_1 e^{j\theta_1}, \alpha_2 e^{j\theta_2}, \cdots, \alpha_N e^{j\theta_N}]$ and the phase shift of RIS is $\theta_N \in [0, 2\pi)$.

Reward: As the formulated optimization problem, the reward function can be established in the form of the optimization objective, which can be expressed as:

$$r = w_d\mathcal{R}_{\text{data}} + w_k\mathcal{R}_{\text{key}} \tag{14}$$

3.4 Computational Complexity

The computational complexity of the proposed algorithm includes training complexity and deployment complexity, which will be analyzed as follows.

Training complexity: Firstly, we calculate the computational complexity of the activation layers. The computational complexity of the ReLU layer is "1", that of the sigmoid layer is "2", and that of the tanh layer is "2". Assume that the total number of nodes in the state normalization layer, ReLU layer, sigmoid layer, and tanh layer are $|\mathcal{S}|, n_r, n_s$, and n_t. Thus, the training complexity for node computation is $O(|\mathcal{S}| + n_r + 2n_s + 2n_t)$. Furthermore, we assume that both the evaluated network and the target network consist of L fully connected layers and the l-th layer has n_l nodes. The training complexity of one for-

ward propagation and two backward propagations can be calculated by $O\left(\sum_{l=0}^{L-1} 3n_l n_{l+1}\right)$. In the PPO algorithm, multiple trajectories need to be sampled from the environment for learning. Supposing that the trajectories sample is N and the length of each trajectory is T, the complexity of the sampling and the update process can be expressed as $O\left(N \cdot T \cdot \left(3\sum_{l=0}^{L-1} n_l n_{l+1}\right)\right)$. The total complexity of the PPO algorithm in the training phase is $O\left(K \cdot N \cdot T \cdot \left(|\mathcal{S}| + n_r + 2n_s + 2n_t + 3\sum_{l=0}^{L-1} n_l n_{l+1}\right)\right)$, where K represents the total number of iterations.

Deployment complexity: Since we only use the policy network π_θ for action selection, sampling and update operations are not involved. Therefore, only the computational complexity of state normalization and one forward propagation needs to be considered. Similar to the above analysis, the complexity of the deployment phase can be expressed as $O\left(\sum_{l=0}^{L-1} n_l n_{l+1}\right) + O(|\mathcal{S}|)$.

4 Simulation Results

In terms of weight factors, the weights of both the data transmission rate w_d and the key generation rate w_k are set to 0.5, which means that the two tasks have equal priority. The learning rate β is set to 0.000 3. This small value ensures that the model parameter updates are relatively stable during the training process, thereby reducing the risk of missing the optimal solution or making the training diverge due to overly large update steps. The discount factor γ is set to 0.99, indicating that the agent places great emphasis on relatively long-term returns. The batch size *batch_size* is set to 64. When parameters are updated each time, 64 samples are extracted from the sample data for calculation. This value can maintain a reasonable computational efficiency while taking into account a certain degree of stability in gradient estimation. In the generalized advantage estimation, the parameter *gae_λ* is set to 0.95, biasing the advantage estimation towards prioritizing the long-term temporal difference error information.

The DQN-based STAG method is proposed to optimize the key generation rate[11]. However, this method makes it difficult to handle continuous action space problems, thereby leading to a dimensional disaster for the large action space or the loss of some action information. The proposed PPO-based STAG method can effectively handle continuous action space and the convergence is stable. Thus, we use the DQN-based method as a benchmark and study the security boundary of the RIS phase shift for STAG with the PPO-based method.

Specifically, the DQN algorithm selects (discrete) phase shift values for the 8 elements in the action space of RIS from $[0, 2\pi)$, and the resolution of the RIS phase-shift is 1 bit. The PPO-based STAG method selects continuous phase shift values for the 8 elements in the continuous action space of RIS from $[0, 2\pi)$. Fig. 2 shows although the DQN-based STAG method converges slightly faster than the PPO-based method, the reward of the former is unstable and not as high as the reward of the latter. The reward of PPO can reach 6.0, while DQN is only 5.7, which has improved the reward by more than 5%. To analyze the optimal solution, we use an exhaustive search optimization method and compare it with the optimization results of the PPO algorithm. In Fig. 2, the optimization results of the PPO algorithm are very close to the optimal result of the exhaustive search optimization, which is demonstrated to be optimally achieved in dynamic wireless environments.

To prove the convergence stability of the PPO-based STAG method in large dimensional action space, we explore the relationship between the reward and the number of RIS reflection units with the continuous phase shift in Fig. 3. When the number of RIS reflection units increases, the key generation rate increases obviously. When the number of RIS reflection units

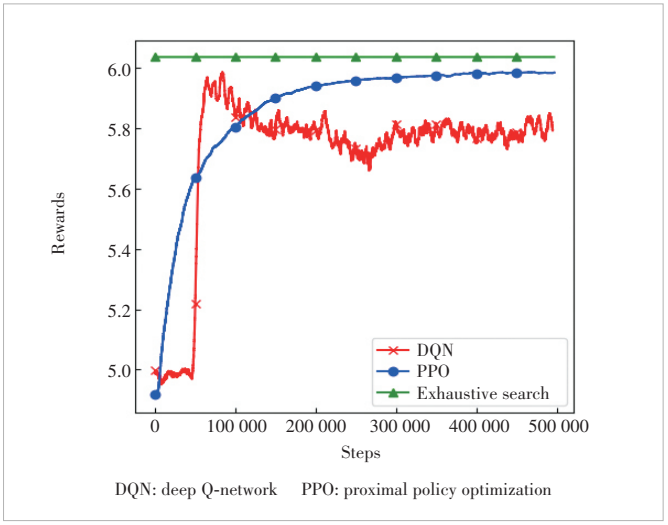


Figure 2. Comparison between DQN algorithm and PPO algorithm

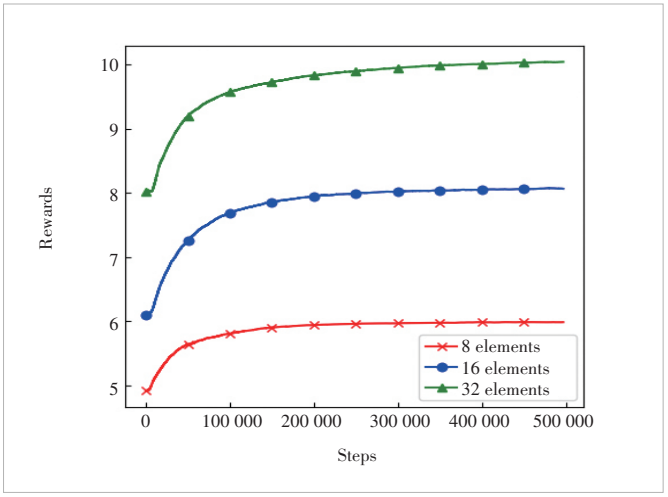


Figure 3. Comparison of different RIS components by using PPO algorithm

is 32, the reward is close to 10, which is twice as much as when the number of RIS reflection units is 8. Specifically, as the number of reflection units rises, the channel gain increases with the assistance of the RIS, thereby improving the STAG performance.

To explore the security boundary of the RIS phase shift and validate the effect of the "one-time pad" with STAG, we study the optimal transmission rate and key generation rate in different weights. Fig. 4 illustrates the relationship between weight and the rate change based on the PPO algorithm. It can be found that as the weight w_k increases, the data transmission rate decreases and the key generation rate increases. The PPO-based STAG method outperforms the DQN-based method both in key generation rate and the data transmission rate. Importantly, the key generation rate and the data generation rate are equal for the proposed PPO-based STAG and the DQN-based STAG when the weight is about 0.675 and 0.92, respectively. It suggests that this weight can achieve "one-time pad" communication via STAG design. Specifically, there is about 90% performance improvement for "one-time pad" communication than that of DQN-based STAG method, which shows the security boundary of the RIS phase shift.

5 Conclusions

In this paper, we study the potential of ICAS to attain perfectly secure communication with the presence of the eavesdropper via the STAG design. Specifically, we consider the dynamic wireless environments and propose a policy gradient algorithm based on PPO, which is to improve the convergence stability of STAG in large-scale action space and explore the security boundary of the RIS phase shift. The simulation results indicate that the proposed PPO-based STAG method has a better performance than the DQN-based STAG method and approaches the optimal exhaustive search, which shows the security boundary of the RIS phase shift. By setting a suitable weight to balance the data transmission rate and communication security, "one-time pad" communication can be achieved.

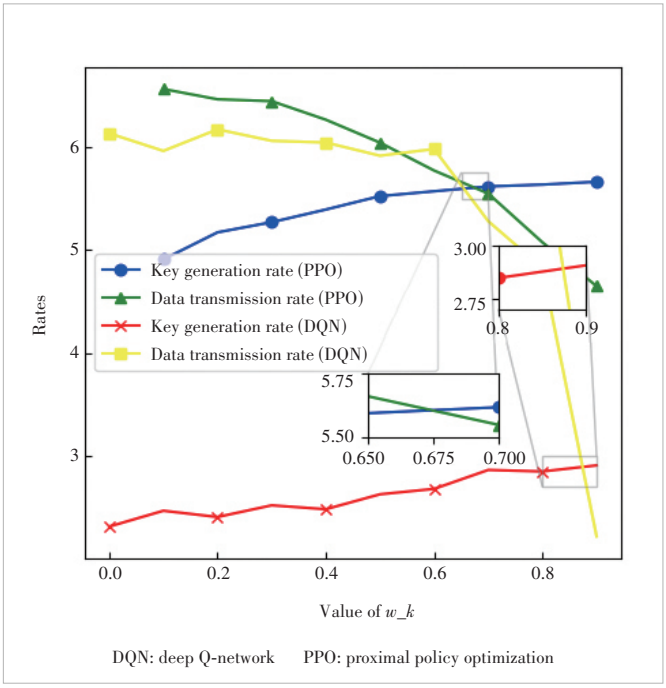


Figure 4. Variation of the data transmission rate and the key generation rate with different weights

References

[1] SANG J, YUAN Y F, TANG W K, et al. Coverage enhancement by deploying RIS in 5G commercial mobile networks: field trials [J]. IEEE wireless communications, 2024, 31(1): 172 – 180. DOI: 10.1109/MWC.011.2200356
[2] GAO N, HAN Y, LI N N, et al. When physical layer key generation meets RIS: opportunities, challenges, and road ahead [J]. IEEE wireless communications, 2024, 31(3): 355 – 361. DOI: 10.1109/MWC.013.2200538
[3] MOARA-NKWE K, SHI Q, LEE G M, et al. A novel physical layer secure key generation and refreshment scheme for wireless sensor networks [J]. IEEE access, 2018, 6: 11374 – 11387. DOI: 10.1109/ACCESS.2018.2806423
[4] MAURER U M. Secret key agreement by public discussion from common information [J]. IEEE transactions on information theory, 1993, 39(3): 733 – 742. DOI: 10.1109/18.256484
[5] PREMNATH S N, JANA S, CROFT J, et al. Secret key extraction from wireless signal strength in real environments [J]. IEEE transactions on mobile computing, 2013, 12(5): 917 – 930. DOI: 10.1109/TMC.2012.63
[6] LI G Y, HU A Q, ZHANG J Q, et al. High-agreement uncorrelated secret key generation based on principal component analysis preprocessing [J]. IEEE transactions on communications, 2018, 66(7): 3022 – 3034. DOI: 10.1109/TCOMM.2018.2814607
[7] JI Z J, YEOH P L, ZHANG D Y, et al. Secret key generation for intelligent reflecting surface assisted wireless communication networks [J]. IEEE transactions on vehicular technology, 2021, 70(1): 1030 – 1034. DOI: 10.1109/TVT.2020.3045728
[8] GU J, OUYANG C J, ZHANG X, et al. RIS-assisted multi-carrier secret key generation in static environments [J]. IEEE wireless communications letters, 2024, 13(10): 2777 – 2781. DOI: 10.1109/LWC.2024.3445268
[9] ZHAO R, QIN Q, XU N Y, et al. SemKey: boosting secret key generation for RIS-assisted semantic communication systems [C]//The 96th Vehicular Technology Conference. IEEE, 2022: 1 – 5. DOI: 10.1109/VTC2022-Fall57202.2022.10013083
[10] XU N Y, NAN G S, TAO X F. Passive eavesdropping can significantly slow down RIS-assisted secret key generation [C]//IEEE Global Communications Conference. IEEE, 2023: 3294 – 3299. DOI: 10.1109/GLOBECOM54140.2023.10437788
[11] GAO N, YAO Y Z, JIN S, et al. Integrated communications and security: RIS-assisted simultaneous transmission and generation of secret keys [J]. IEEE transactions on information forensics and security, 2024, 19: 7573 – 7587. DOI: 10.1109/TIFS.2024.3436885
[12] TANG J W, XU S H, YANG F, et al. Recent developments of transmissive reconfigurable intelligent surfaces: a review [J]. ZTE Communications, 2022, 20(1): 21 – 27. DOI: 10.12142/ZTECOM.202201004
[13] LIU Y W, LIU X, MU X D, et al. Reconfigurable intelligent surfaces: principles and opportunities [J]. IEEE communications surveys & tutorials, 2021, 23(3): 1546 – 1577. DOI: 10.1109/COMST.2021.3077737
[14] GAO N, QIN Z J, JING X J, et al. Anti-intelligent UAV jamming strategy via deep Q-networks [J]. IEEE transactions on communications, 2020, 68(1): 569 – 581. DOI: 10.1109/TCOMM.2019.2947918
[15] LUONG N C, HOANG D T, GONG S M, et al. Applications of deep rein-

forcement learning in communications and networking: a survey [J]. IEEE communications surveys and tutorials, 2019, 21(4): 3133 – 3174
[16] QIAN X, DI RENZO M, LIU J, et al. Beamforming through reconfigurable intelligent surfaces in single-user MIMO systems: SNR distribution and scaling laws in the presence of channel fading and phase noise [J]. IEEE wireless communications letters, 2021, 10(1): 77 – 81. DOI: 10.1109/LWC.2020.3021058
[17] ZHANG H Q, LI X, GAO N, et al. A deep reinforcement learning approach to two-timescale transmission for RIS-aided multiuser MISO systems [J]. IEEE wireless communications letters, 2023, 12(8): 1444 – 1448. DOI: 10.1109/LWC.2023.3278171
[18] LU T Y, CHEN L Q, ZHANG J Q, et al. Joint precoding and phase shift design in reconfigurable intelligent surfaces-assisted secret key generation [J]. IEEE transactions on information forensics and security, 2023, 18: 3251 – 3266. DOI: 10.1109/TIFS.2023.326888

Biographies

FAN Kaiqing received his BS degree in computer science from Nanjing University of Finance and Economics, China in 2023. He is currently pursuing his MS degree with the School of Cyber Science and Engineering, Southeast University, China. His research interests are RIS-assisted physical layer security and deep reinforcement learning.

YAO Yuze received his BS degree in information security from China University of Mining and Technology, China in 2023. He is currently pursuing his MS degree with the School of Cyber Science and Engineering, Southeast University, China. His research interests include wireless communication security and deep reinforcement learning.

GAO Ning (ninggao@seu.edu.cn) received his PhD degree in information and communications engineering from Beijing University of Posts and Telecommunications, China in 2019. From 2017 to 2018, he was a visiting PhD student with the School of Computing and Communications, Lancaster University, UK. From 2019 to 2022, he was a research fellow with the National Mobile Communications Research Laboratory, Southeast University, China. He is currently an associate professor with the School of Cyber Science and Engineering, Southeast University. His research interests include AI enabled wireless communications and security, reconfigurable intelligent surfaces (RIS), and UAV communications.

LI Xiao received her PhD degree in communication and information systems from Southeast University, China in 2010. Then, she joined the School of Information Science and Engineering, Southeast University, where she has been a professor of information systems and communications since July 2020. From January 2013 to January 2014, she was a postdoctoral fellow at The University of Texas at Austin, USA. Her current research interests include massive MIMO, reconfigurable intelligent surface assisted communications, and intelligent communications. She was a recipient of the 2013 National Excellent Doctoral Dissertation of China for her PhD dissertation.

JIN Shi received his PhD degree in communications and information systems from Southeast University, China in 2007. From June 2007 to October 2009, he was a research fellow with the Adastral Park Research Campus, University College London, UK. He is currently a faculty member with the National Mobile Communications Research Laboratory, Southeast University. His research interests include wireless communications, random matrix theory, and information theory. He was an associate editor of *IEEE Transactions on Wireless Communications*, *IEEE Communications letters*, and *IET Communications*. He serves as an area editor of *IEEE Transactions on Communications* and *IET Electronics Letters*.

Endogenous Security Through AI-Driven Physical-Layer Authentication for Future 6G Networks

MENG Rui[1], FAN Dayu[1], XU Xiaodong[1,2], LYU Suyu[3], TAO Xiaofeng[4]

(1. State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. Department of Broadband Communication, Peng Cheng Laboratory, Shenzhen 518066, China;
3. School of Information Science and Technology, Beijing University of Technology, Beijing 100124, China;
4. National Engineering Laboratory for Mobile Network Technologies, Beijing University of Posts and Telecommunications, Beijing 100876, China)

DOI: 10.12142/ZTECOM.202501004

https://kns.cnki.net/kcms/detail/34.1294.TN.20250210.1626.004.html, published online February 11, 2025

Manuscript received: 2025-01-09

Abstract: To ensure the access security of 6G, physical-layer authentication (PLA) leverages the randomness and space-time-frequency uniqueness of the channel to provide unique identity signatures for transmitters. Furthermore, the introduction of artificial intelligence (AI) facilitates the learning of the distribution characteristics of channel fingerprints, effectively addressing the uncertainties and unknown dynamic challenges in wireless link modeling. This paper reviews representative AI-enabled PLA schemes and proposes a graph neural network (GNN)-based PLA approach in response to the challenges existing methods face in identifying mobile users. Simulation results demonstrate that the proposed method outperforms six baseline schemes in terms of authentication accuracy. Furthermore, this paper outlines the future development directions of PLA.

Keywords: physical-layer authentication; artificial intelligence; wireless security; intelligent authentication

Citation (Format 1): MENG R, FAN D Y, XU X D, et al. Endogenous security through AI-driven physical-layer authentication for future 6G networks [J]. ZTE Communications, 2025, 23(1): 18 - 29. DOI: 10.12142/ZTECOM.202501004

Citation (Format 2): R. Meng, D. Y. Fan, X. D. Xu, et al., "Endogenous security through AI-driven physical-layer authentication for future 6G networks," *ZTE Communications*, vol. 23, no. 1, pp. 18 - 29, Mar. 2025. doi: 10.12142/ZTECOM.202501004.

1 Introduction

1.1 Background

To further accelerate the realization of the Internet of Everything, 6G mobile networks will integrate a multitude of enabling technologies, with a goal of achieving extensive coverage, high bandwidth, low latency, and highly reliable communications[1]. Currently, the official launch of the first 6G standard project by the 3rd Generation Partnership Project (3GPP) marks the transition of 6G from technical pre-research to the standardization phase, signaling the start of a critical period for blueprint formulation. However, as an emerging technology, 6G will introduce more complex security challenges[2 - 3]. The future three-dimensional and fully integrated communication network, characterized by diverse, resilient, and distributed topologies, involves numerous heterogeneous nodes, dynamic resource management, and ubiquitous diverse connections, thereby increasing network complexity and security risks[4 - 5]. While various enabling technologies offer numerous potential advantages and application prospects, they also introduce certain security problems[6]. For example, attackers can exploit user interference caused by a vast number of antennas and devices in ultra-massive multi-input multi-output (UM-MIMO) systems to eavesdrop on and tamper with data. To ensure secure communication and transmission, threat detection and defense, and data confidentiality and integrity in 6G networks, it is crucial to redesign security safeguard mechanisms to achieve intelligent, flexible, and real-time endogenous security.

1.2 Physical-Layer Authentication

As a complement to traditional upper-layer authentication protocols, physical-layer authentication (PLA), with its high reliability, lightweight design, and exceptional compatibility, is considered an endogenous security protection strategy[7]. Primarily, the characteristics of physical-layer attributes, based on the inherent randomness of channels and the uniqueness of space-time-frequency, which are closely related to communication links, devices, and locations, can represent unique

identity signatures for legitimate users, making it extremely difficult for attackers to extract, imitate, or forge them[8]. Secondly, PLA cleverly bypasses high-level signaling processes, allowing its access points to obtain the channel state information (CSI) of legitimate users during the channel estimation phase, significantly reducing computational resource consumption[9]. Furthermore, even if incompatible devices may face obstacles in decoding each other's upper-layer signaling, they can still successfully parse the bit stream at the physical-layer, further broadening the application and flexibility[10].

Recently, a growing number of researchers have designed artificial intelligence (AI)-empowered PLA methods to effectively address the uncertainty and unknown dynamic challenges in wireless link modeling[11]. Advanced machine learning (ML) algorithms can intelligently learn the distribution characteristics of channel fingerprints and optimize the authentication threshold in dynamic environments, achieving adaptive online authentication[12]. Additionally, unsupervised learning algorithms help build a malicious node detection model without prior knowledge of the attacker's location or attack frequency[13]. Furthermore, deep learning (DL) technology excels at learning high-dimensional fingerprint features and classifying a large number of samples, enabling the identification of large-scale or even ultra-large-scale devices[14]. In summary, compared with traditional PLA methods, AI-empowered PLA has several advantages. It overcomes the challenges of modeling the uncertainty and unknown dynamics of wireless links, achieves adaptive threshold authentication, possesses greater universality without needing extensive prior information, exhibits higher scalability, and is capable of identifying ultra-large-scale equipment[15].

1.3 Contributions

The main contributions of this paper are summarized as follows.

1) We review representative AI-based PLA research, which is classified into radio frequency (RF) fingerprint extraction, fingerprint data augmentation, lightweight authentication models, authentication parameter optimization, multi-attacker identification, and physical-layer key generation for frequency-division duplexing (FDD) systems.

2) We propose a graph neural network (GNN)-based PLA scheme to identify mobile multiusers. Unlike most existing convolutional neural network (CNN)-based PLA schemes, the proposed scheme can learn the spatial correlation among various CSI fingerprint dimensions introduced by reconfigurable intelligent surfaces (RISs) through modeling the nodes and edges. Furthermore, the scheme also captures the temporal correlation between fingerprints and within fingerprint sequences through dynamic graphs and temporal convolution learning. The simulations demonstrate the superiority of the proposed scheme over six baseline schemes.

3) We envision the future research direction of intelligent PLA for 6G, including semantic fingerprint-based PLA, large AI model-based PLA, cross-layer PLA, multi-modal signature-based PLA, distributed autonomous PLA, and PLA for emerging applications.

2 Existing AI-Enabled PLA Approaches

In Table 1, we provide a brief review of existing AI-empowered PLA schemes, which is explained in detail below.

2.1 RF Fingerprint Extraction

The extraction of RF fingerprints relies on the hardware variations of transmitters, such as digital-to-analog converters (DAC), in-phase/quadrature (I/Q) modulators, and power amplifiers. These differences result in distinct inherent properties among radiation sources of the same model and batch. Traditional extraction methods often depend on preprocessing techniques, such as time synchronization and phase offset compensation, as well as expert feature transformation meth-

Table 1. Brief review on existing AI-empowered PLA schemes

Categories	Motivations	Methods	Performance
RF fingerprint extraction	The extraction of RF fingerprints requires much prior information	CNN[16], RNN[17], attention mechanism[18], and CVNN[19]	Realizing end-to-end RF fingerprint extraction
Fingerprint data augmentation	Insufficient fingerprint samples lead to overfitting issues of PLA models, thus limiting authentication performance	Added noise-based[20] and generated fingerprint-based[21] schemes	Enhancing the generalization of PLA models
Lightweight authentication model	To identify ultra-large-scale devices, PLA models usually have a large number of parameters and deep structures	Transfer learning-based[22] and network compression-based[23] schemes	Reducing the deployment complexity of PLA models
Authentication parameter optimization	Optimizing detection thresholds is challenging in complex channel environments	RL[24-25]	Achieving the automatic optimization of authentication parameters
Multi-attacker identification	The prior information of multi-attackers is difficult to obtain in actual applications	Clustering[13], OCC[26], and GMM[27]	Realizing authentication without knowing the prior information of attackers
Physical-layer key generation for FDD systems	In FDD systems, uplink and downlink transmissions work in different frequency bands, and their channel frequency responses are no longer reciprocal	Generative AI[28]	Improving the key generation ratio

AI: artificial intelligence
CNN: convolutional neural network
CVNN: complex-valued neural network
FDD: frequency-division duplexing
GMM: Gaussian mixture model
OCC: one class classification
PLA: physical-layer authentication
RL: reinforcement learning
RNN: recurrent neural network

ods like the short-time Fourier transform and wavelet transform. However, these processes require prior information, limiting the practical applicability. In recent years, with the advantages of DL in feature extraction, the acquisition of RF fingerprints gradually overcomes the dependence on prior information and manually optimizing parameters, and only requires preprocessing processes such as normalization and interpolation. DL is realized by neural networks, such as CNN[16], recurrent neural networks (RNN)[17], attention mechanisms[18], and complex-valued neural networks (CVNN)[19].

Specifically, Ref. [16] presents a novel DL-based RF fingerprint identification approach to IoT terminal authentication, leveraging the differential constellation trace figure (DCTF) to extract RF fingerprint features without synchronization. CNN is designed to identify devices using DCTF features. It offers high accuracy, requires no prior information, and maintains low complexity. Ref. [17] explores RNNs for autonomous wireless system deployments in RF environments. By utilizing the temporal properties of received radio signals, Ref. [17] proposes a transmitter fingerprinting technique for device identification. Ref. [17] implements three RNN models, namely Long Short-Term Memory (LSTM), the Gated Recurrent Unit (GRU), and ConvLSTM, using I/Q time series data collected from eight universal software radio peripheral (USRP) software defined radio (SDR) transmitters. By exploiting temporal variations and spatial dependencies in the data, the model learns unique feature representations for transmitter identification. Ref. [18] presents a novel multi-channel attentive feature fusion method for RF fingerprinting. Unlike other models that rely on a single representation of radio signals, the proposed method integrates multiple representations, such as in-phase and quadrature samples, carrier frequency offsets, and frequency transform coefficients. By employing a shared attention module, Ref. [18] adaptively fuses neural features extracted from these different channels, optimizing their weights during training. Additionally, a convolution-based ResNeXt block is implemented to map the fused features to specific device identities. Given that wireless signal information is encoded in complex basebands, Ref. [19] studies the application of CVNNs to develop device fingerprints through supervised learning.

2.2 Fingerprint Data Augmentation

The training of DL-based PLA models usually requires a large number of fingerprint samples. However, it is challenging to obtain sufficient fingerprint samples in practical applications. To address this issue, data augmentation is an effective approach to enhancing the model generalization and improving the authentication accuracy. We divide the existing fingerprint data augmentation schemes into two subcategories: added noise-based[20] and generated fingerprint-based[22] schemes. The former employs Gaussian noises to mitigate model overfitting, while the latter enhances sample richness by generating additional fingerprint samples.

Specifically, Ref. [20] aims to enhance authentication performance with minimal training data by applying Gaussian noises in a smooth latent space, thus improving generalization and interpretability. The proposed scheme avoids reliance on synthetic samples while providing insights into the authentication process through the defined Fingerprint Library. This allows for a better understanding of how input channel impulse responses (CIRs) correlate with authentication outcomes. Ref. [21] employs three data augmentation algorithms to expedite the model establishment and improve authentication success rates. By integrating deep neural networks with these augmentation methods, the scheme not only enhances performance but also accelerates training, even with limited samples.

2.3 Lightweight Authentication Model

To realize the identification of ultra-large-scale devices, authentication models typically possess a large number of parameters and deep structures to learn multi-level and abstract fingerprint features. To reduce the computation and storage requirements of the PLA model without sacrificing most performance, researchers have designed transfer learning-based[22] and network compression-based[21] PLA schemes. The former can quickly identify the physical-layer fingerprints of different equipment types in unknown radio environments with only a few training samples through a pre-trained model[22]. For example, Ref. [22] introduces transfer learning to realize swift online user authentication, crucial for latency-sensitive applications like edge computing. The latter employs lightweight technologies, such as quantization, grouping convolution, and distillation, to reduce the parameters and calculation of PLA models. For instance, Ref. [23] introduces network compression techniques to reduce the model complexity and size. Despite the high model complexity and size of CVNNs, the proposed approach ensures satisfactory identification performance.

2.4 Authentication Parameter Optimization

PLA is typically modeled as a hypothesis testing problem, where the authentication result is obtained by comparing the difference between the signal to be authenticated and a reference signal with a detection threshold. Therefore, optimizing the detection threshold is crucial for authentication performance. Due to complex multipath effects, time-varying characteristics of channels, noise interference, and other factors, deriving the detection threshold becomes increasingly difficult. To address this issue, RL, through continuous interaction with the environment, can learn how to make optimal authentication decisions without fully understanding the channel model. Ref. [24] frames the interactions between a legitimate receiver and spoofers as a zero-sum authentication game. The receiver adjusts its test threshold to maximize utility based on the Bayesian risk in spoofing detection, while spoofers aim to minimize this utility by varying their attack frequencies. Since obtaining precise channel parameters beforehand is challeng-

ing, Ref. [24] introduces spoofing detection schemes based on Q-learning and Dyna-Q. These schemes leverage RL to determine the optimal test threshold for spoofing detection. Ref. [25] presents a novel controller area network (CAN) bus authentication framework designed to protect message exchanges against spoofing attacks. The proposed framework leverages RL to optimize the selection of authentication modes and parameters. By implementing the Dyna architecture with the double estimator, the framework enhances authentication accuracy without necessitating changes to the CAN bus protocol or electronic control unit components.

2.5 Multi-Attacker Identification

For detection attack scenarios, a suitable assumption is that the attackers' prior information is unknown, and often multi-attackers are present to confuse legitimate receivers. To address this challenge, unsupervised learning can construct an authentication model without requiring the attackers' prior information or training fingerprint set. By establishing decision boundaries, the detection of multi-attackers is achieved. Ref. [13] proposes a multi-attribute-based approach that considers the inherent correlation among physical-layer attributes. To manage the exponential computational complexity of correlated analysis, Ref. [13] introduces a reconstruction and heuristic algorithm to find a suboptimal solution with reduced complexity. An unsupervised machine learning-based non-parametric clustering algorithm is proposed to enhance authentication reliability. The proposed approach does not require prior information or a training set, thereby improving its universality. Ref. [26] assesses and compares the performance of various approaches under different channel conditions. Ref. [26] evaluates statistical decision methods and ML classification techniques, including one-class classifiers for scenarios with no forged messages or conventional binary classifiers when forged messages are present. Numerical results demonstrate that one-class classification algorithms achieve the lowest missed detection probability under low spatial correlation. Ref. [27] utilizes GMMs to identify spoofing attackers by clustering messages based on probabilistic models of different transmitters. A 2D feature measure space is used to preprocess channel information, and a pseudo adversary model is developed to enhance detection performance against spoofers operating through unknown channels.

2.6 Physical-Layer Key Generation for FDD Systems

Physical-layer key generation offers a robust and efficient method for secure key generation by leveraging the unique properties of wireless channels. Exploiting the reciprocity and time-varying nature of these channels ensures that both communicating parties can generate identical keys with minimal communication overhead and hardware requirements. The implementation of physical-layer key generation relies on the reciprocity of channels. However, in FDD systems, the uplink (from a user to a base station) and downlink (from a base station to a user) operate on separate frequency bands. This duplexing method allows for simultaneous uplink and downlink communications, but it also introduces a frequency difference. The properties of the wireless channel, such as path loss, shadowing, and multipath effects, are functions of frequency. Consequently, the frequency difference disrupts the channel reciprocity. To address this issue, generative AI is a promising approach. Ref. [28] introduces a novel physical-layer key generation scheme for FDD systems, addressing the challenges of extracting common features in non-reciprocal channels, and employs DL to create a feature mapping function between different frequency bands, enabling two users to generate highly similar channel features. Ref. [28] also proves the existence of a band feature mapping function using a feedforward network with a single hidden layer and proposes a key generation neural network for reciprocal channel feature construction.

3 Proposed PLA Scheme for Mobile Users

This section provides the GNN-based PLA to identify mobile users, including the research motivation, networks and channel models, problem formulation, research methods, and simulation results.

3.1 Motivation

The accuracy and reliability of CSI fingerprints are crucial for PLA. However, their quality is often constrained in some scenarios such as the Industrial Internet of Things (IIoT) due to multipath fading, obstacle interferences, and complex electromagnetic environments. To tackle this issue, RIS intelligently adjusts the wireless propagation environment, significantly boosting the expected signal power at the receiver[29]. Nevertheless, existing CNN-based PLA models frequently overlook the potential interdependencies among various CSI dimensions. With the integration of RIS, the wireless environment has transformed, resulting in a strong correlation among diverse dimensional features of CSI fingerprints. Hence, the primary challenge lies in fully extracting the intrinsic features of these reconfigurable channel fingerprints.

Furthermore, in certain scenarios, smart devices are frequent in motion. For example, mobile terminals in logistics and production lines augment efficiency and flexibility, while unmanned vehicles and mobile robots engaged in data collection and monitoring tasks enhance real-time analysis and decision-making capabilities. Since CSI is a location-specific physical-layer attribute, user movement alters the distribution of CSI, with greater deviations as the distance from the transmitter increases[30]. Consequently, leveraging CSI-based PLA methods to identify mobile users poses another significant challenge.

To address the first challenge, we deployed GNNs to capture the dependencies and topological structures among various CSI dimensions introduced by the RIS. Existing CNN-based PLA models frequently neglect the underlying depen-

dency relationships among different CSI dimensions. In addition, RNNs have certain limitations in handling sequence data, particularly long sequences, which restricts their ability to capture long-term dependencies. In contrast, GNNs, through the connections of nodes and edges, can naturally capture the correlations among multi-dimensional channel features. These direct or indirect correlations are transmitted through paths between nodes. For example, Ref. [31] models MIMO CSI prediction as a multivariate time-series forecasting problem and introduces GNNs to exploit both spectral and temporal correlations between historical and future CSI.

To tackle the second issue, we formulated the variations of CSI fingerprints in mobile scenarios as time series. We then integrated temporal convolution networks and dynamic GNNs to fully exploit the temporal correlations both among CSI samples and within sequences of CSI samples. Unlike static GNNs, dynamic GNNs can capture both spatial and temporal dependencies among variables and excel at processing multivariate time series data.

3.2 Network Model

As depicted in Fig. 1, we consider a multiuser access authentication scenario, wherein K users engage in communication with the receiver (Bob) across distinct time slots. Given that users are in constant motion, the distance between them is assumed to exceed half a wavelength, ensuring the uniqueness of their fingerprints. To bolster signal strength and broaden coverage, RISs are utilized to redirect the incident signal toward the target area by adjusting the reflected signal. This enhances the quality of channel fingerprints in areas affected by signal blind spots or weak signal reception. Notably, RISs are controlled by Bob. Additionally, edge servers stationed at Bob's location are leveraged to optimize the deployment performance of AI-driven PLA models.

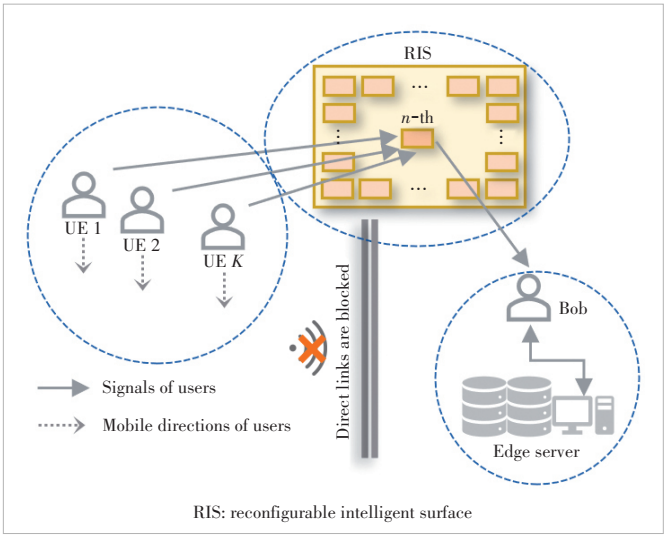


Figure 1. System model of a multiuser access authentication scenario

3.3 Channel Model

N_T and N_R represent the numbers of antennas of each user and of Bob, and the received signal at Bob can be denoted as:

$$\boldsymbol{Y}_S = \boldsymbol{Q}\boldsymbol{X}_S + \boldsymbol{W} \tag{1}$$

where $\boldsymbol{X}_S$ with N_T-size column denotes the transmitted signal, and $\boldsymbol{W} \sim \mathcal{CN}(0,\boldsymbol{\sigma}^2)$ with N_R-size column denotes Gaussian noises. $\boldsymbol{Q} = \boldsymbol{H}\boldsymbol{\Psi}\boldsymbol{G} \in \mathbb{C}^{N_R \times N_T}$ represents the hierarchical channel matrix from the user to Bob through RISs, where $\boldsymbol{H} \in \mathbb{C}^{N_R \times N}$ and $\boldsymbol{G} \in \mathbb{C}^{N \times N_T}$ respectively stand for the channel matrices from RISs to Bob and from the user to RISs, and $\boldsymbol{\Psi} = \mathrm{diag}\left(\boldsymbol{\psi}_0,\cdots,\boldsymbol{\psi}_{N-1}\right) \in \mathbb{C}^{N \times N}$ represents the response matrix of RISs with N denoting the number of elements of RISs. $\boldsymbol{\psi}_n = A_n\left(\theta_n\right)e^{j\theta_n}$ with $A_n\left(\theta_n\right)$ and $e^{j\theta_n}$ respectively denoting the controllable magnitude and phase response of the n-th RIS element. $\boldsymbol{H}$ and $\boldsymbol{G}$ are modeled as Rician channels, which are denoted as:

$$\boldsymbol{H} = \sqrt{\frac{PL\kappa_H}{1+\kappa_H}}\bar{\boldsymbol{H}} + \sqrt{\frac{PL}{1+\kappa_H}}\tilde{\boldsymbol{H}} \tag{2}$$

and

$$\boldsymbol{G} = \sqrt{\frac{PL\kappa_G}{1+\kappa_G}}\bar{\boldsymbol{G}} + \sqrt{\frac{PL}{1+\kappa_G}}\tilde{\boldsymbol{G}} \tag{3}$$

where $\bar{\boldsymbol{H}}$ and $\bar{\boldsymbol{G}}$ represent line of sight (LoS) paths, κ_H and κ_G represent Rician factors, and $\tilde{\boldsymbol{H}}$ and $\tilde{\boldsymbol{G}}$ denote non-LoS (NLoS) paths. PL represents the corresponding path loss. The configurable fingerprints $\boldsymbol{x}$ are acquired via channel estimation, which is not the focus of this paper and can be accomplished through various techniques, such as compressed sensing, matrix factorization, and DL methods[32].

3.4 Problem Formulation

Due to the multidimensional nature of complex CSI fingerprints in mobile scenarios, these fingerprints can be represented as multivariate time series $\boldsymbol{X} = \{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_d\} \in \mathbb{R}^{d \times l}$, where $d = 2N_RN_T$ signifies the dimension of CSI fingerprints. Each time series component can be denoted as $\boldsymbol{x}_i = \{\boldsymbol{x}_{i,1},\boldsymbol{x}_{i,2},\cdots,\boldsymbol{x}_{i,l}\}$, where $i = 1,2,\cdots,d$ and $l \in \mathbb{N}^*$ denotes the length of CSI fingerprint sequences. The authentication problem is formulated as a classification task from $\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_m\}$ to $\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_m\}$, aiming to predict the identity $\boldsymbol{y}$ of the CSI fingerprint sequence $\boldsymbol{X}$. Here, $\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_m\}$ corresponds to the identity labels of the CSI fingerprint sequences $\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_m\} \in \mathbb{R}^{m \times d \times l}$, with m denoting the number of CSI fingerprint sequences.

3.5 Proposed GNN-Based PLA Scheme

As illustrated in Fig. 2, the proposed PLA scheme includes training and authentication stages. Fig. 3 illustrates the de-

tailed training process, including fingerprint acquisition, fingerprint preprocessing, graph initialization, temporal convolutional networks, dynamic GNN, hierarchical pooling, and authentication result output modules.

3.5.1 Fingerprint Acquisition

As described in Section 3.3, the cascade CSI fingerprints can be acquired through channel estimation. In this paper, artificial noise is considered to verify the authentication performance versus different signal-to-noise ratio (SNR) conditions.

3.5.2 Fingerprint Preprocessing

The training CSI dataset is composed of CSI fingerprints and corresponding identity labels, which are represented as:

$$\boldsymbol{X}_{\text{train}} = \left[\underbrace{\boldsymbol{X}_1^1,\cdots,\boldsymbol{X}_1^{N_1}}_{N_1}, \underbrace{\boldsymbol{X}_2^1,\cdots,\boldsymbol{X}_2^{N_2}}_{N_2},\cdots, \underbrace{\boldsymbol{X}_K^1,\cdots,\boldsymbol{X}_K^{N_K}}_{N_K} \right] \tag{4},$$

$$\boldsymbol{Y}_{\text{train}} = \left[\underbrace{\boldsymbol{L}_1,\cdots,\boldsymbol{L}_1}_{N_1}, \underbrace{\boldsymbol{L}_2,\cdots,\boldsymbol{L}_2}_{N_2},\cdots, \underbrace{\boldsymbol{L}_K,\cdots,\boldsymbol{L}_K}_{N_K} \right] \tag{5},$$

where N_k denotes the number of CSI sequences of the k-th user, $k \in [1,K]$, and $\boldsymbol{L}_k$ represents the corresponding identity label encoded by one-hot coding[33].

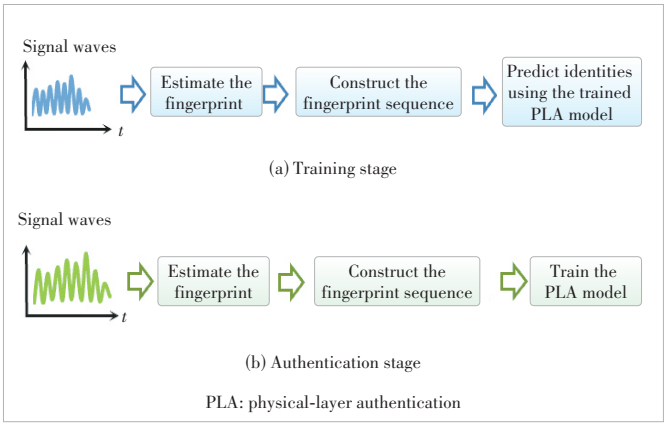


Figure 2. Proposed PLA approach

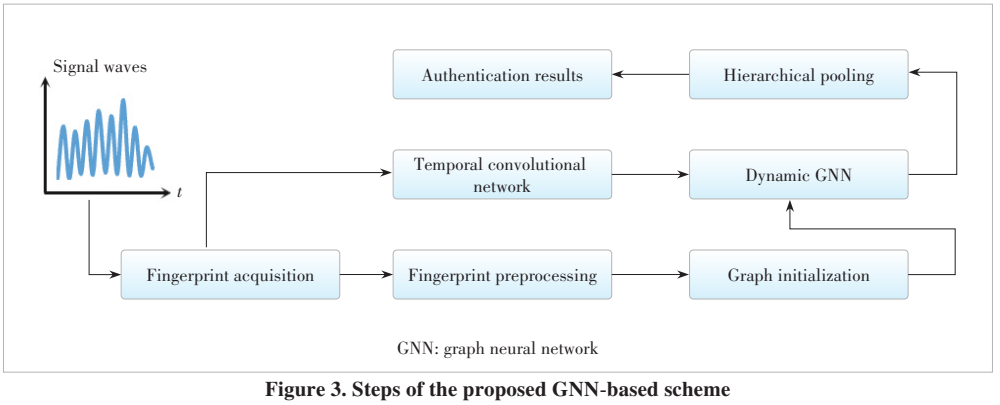


Figure 3. Steps of the proposed GNN-based scheme

3.5.3 Graph Initialization

Nodes and edges collectively form the core structure of a graph, typically denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$[34]. Nodes $\mathcal{V}$, serving as the fundamental building blocks of a graph, represent entities or objects within the graph, specifically the CSI fingerprint sequences of users. Edges $\mathcal{E}$ play the pivotal role of bridges connecting nodes, revealing the correlations and interactions among them. Edges $\mathcal{E}$ can be either directed or undirected, and may even be assigned weights to quantify the strength or importance of the relationships between nodes $\mathcal{V}$.

The essence of GNNs lies in deeply extracting the representations of nodes and edges. Through continuous learning and updating of node features, more enriched and insightful node representations can be generated. Leveraging the connectivity among nodes and the characteristic information of edges, operations such as message passing and graph structure learning are conducted, further extracting the global features of the graph.

The relationships between various nodes are represented through adjacency matrices, where each node is assigned two values representing the source node and the target node[35]. Consequently, each time series corresponds to two vectors, $\boldsymbol{\lambda}$ and $\boldsymbol{\varphi}$, both with the length of d. The values of $\boldsymbol{\lambda}$ and $\boldsymbol{\varphi}$ are randomly initialized. The adjacency matrix can be expressed as:

$$\boldsymbol{A} = \boldsymbol{\lambda}^T \cdot \boldsymbol{\varphi} \tag{6}.$$

Furthermore, we set most of the adjacency matrix's elements to zero, thereby rendering it sparser and reducing the number of elements that need to be computed. Specifically, for the adjacency matrix of each time series, only the top k elements with the highest weights are retained, while the other values are set to zero.

3.5.4 Temporal Convolutional Network

Temporal convolutional networks focus on capturing the temporal dependencies within each dimension of the CSI fingerprint by utilizing three CNN layers with different convolutional kernels, and applying padding operations to ensure that the output length matches the input CSI fingerprint sequence[36]. As illustrated in Fig. 4, in CNNs, neurons deviate from the fully connected architecture of traditional neural networks by adopting a locally connected approach. Specifically, each neuron establishes a connection to a local region of the input data, known as the recep-

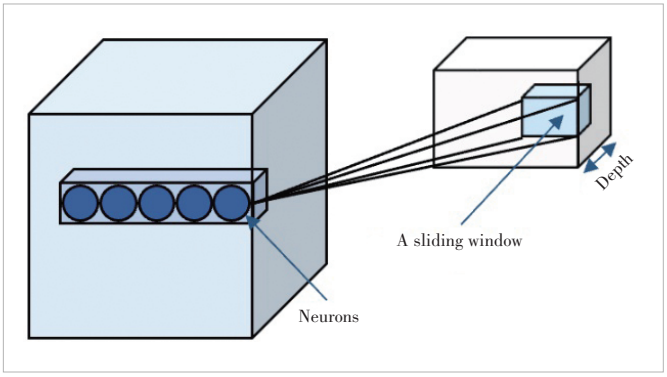


Figure 4. Representation of the convolution operation in CNN layers

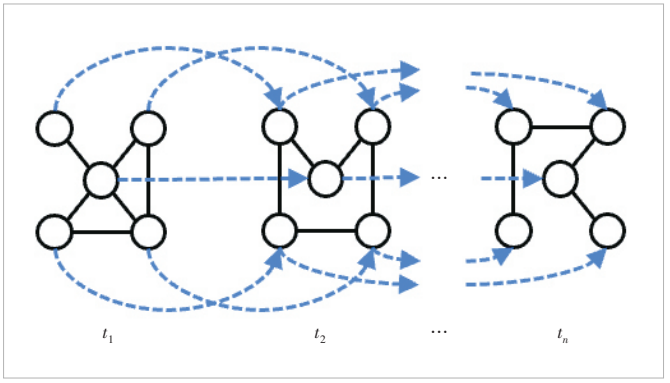


Figure 5. Dynamic graph

tive field, via a convolution kernel (often implemented as a window function). Typically, the depth of the convolution kernel aligns with the depth of the input data. Each convolution kernel is designed to generate a feature map, meaning that multiple convolution kernels collectively yield multiple feature maps, contributing to the depth of the output data.

The learned characteristics of the l-th CNN layer can be denoted as:

$$\boldsymbol{X}_l = \sigma\left(\boldsymbol{W}_l * \boldsymbol{X}_{l-1} + \boldsymbol{B}_l\right) \tag{7},$$

where $\boldsymbol{X}_l$ serves as both the output from the $(l-1)$-th CNN layer and the input to the l-th CNN layer, σ represents the activation function and * denotes the convolution operation. Additionally, $\boldsymbol{W}_l$ and $\boldsymbol{B}_l$ represent the weight and bias matrices, respectively, within the l-th CNN layer.

3.5.5 Dynamic GNN

GNNs are broadly classified into static and dynamic graph categories. Static graphs are particularly suited for scenarios featuring unchanging topological structures, such as user relationship graphs in social networks. Conversely, dynamic graphs excel in managing evolving graph structures and attributes, akin to traffic networks where vehicle positions vary over time[37]. In mobile wireless communication scenarios, shifts in user positions result in continuous alterations in the distribution of CSI fingerprints. Consequently, dynamic graphs are employed to capture the temporal dynamics inherent in CSI fingerprint sequences.

As shown in Fig. 5, for all graphs except the first one, an identical number of vertices are added to represent the CSI fingerprint characteristics of the corresponding vertices from the previous time series. Directed edges are assigned between vertices from the previous time window $v_{(t-1,n)}$ and the current time series $v_{(t,n)}$ to establish associations.

3.5.6 Hierarchical Pooling

By combining graph pooling and temporal processing, this module utilizes hierarchical pooling to decrease the number of nodes, thereby circumventing the information loss inherent in techniques like max pooling and average pooling[38]. As shown in Fig. 6, at each hierarchical level, nodes are converged through temporal convolutions to extract temporal features, and the adjacency matrix is then updated using convolutional weights.

3.5.7 Authentication Results

This module averages the values in the feature graph through average pooling to obtain a fixed-length vector. This vector is then mapped to a logic vector through a fully connected layer, and finally, the authentication result is obtained through the softmax function.

3.6 Simulation Results and Analysis

3.6.1 Baseline Schemes

We consider six baseline schemes as follows.

• K-nearest neighbor (KNN)[39]: Given a test sample, KNN searches for the k nearest fingerprint samples (neighbors) in the training dataset. Based on the information of these k neighbors, the identity of the test fingerprint sample is predicted.

• Naive Bayes (NB)[40]: NB assumes that the features are conditionally independent of each other given the identity label. Based on this assumption and Bayes' theorem, it calculates the posterior probability of each class for a given sample and assigns the sample to the class with the highest posterior probability.

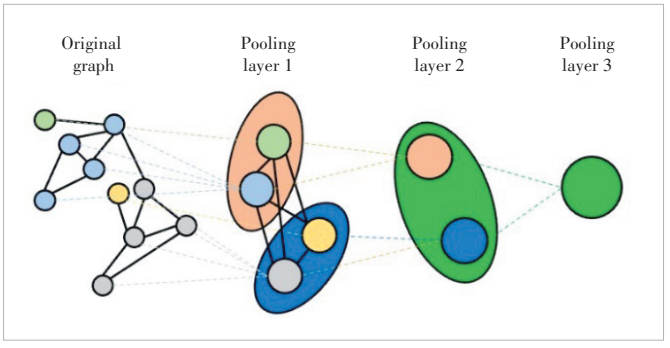


Figure 6. Hierarchical pooling

• Gradient boosting decision tree (GBDT)[39]: GBDT iteratively constructs multiple decision trees and minimizes the loss function through gradient descent, thereby gradually improving prediction accuracy. Its core idea is to build a strong learner using weak learners. In each iteration, GBDT adds a new decision tree to the current model to fit the residuals between the predictions of the previous model and the true values, thereby progressively refining the identity predictions.

• Regularized gradient boosting optimization (RGBO) [30]: Compared with GBDT, RGBO utilizes a second-order Taylor expansion to approximate the changes of the loss function, enabling it to more accurately estimate the descent direction at each iteration, thereby accelerating convergence speed and improving prediction accuracy. Additionally, RGBO incorporates a regularization term into the objective function to control the complexity of the model and prevent overfitting.

• Improved gradient boosting optimization (IGBO)[30]: Unlike RGBO, IGBO efficiently processes data, reduces memory consumption, and enhances training speed by optimizing the sampling process of fingerprints.

• Hybrid method (combining CNNs and RNNs)[41]: CNNs excel at feature extraction from static data, particularly in isolating local features within images. Conversely, RNNs are adept at handling the dependencies inherent in time series data, effectively retaining and utilizing past information. Consequently, the hybrid method merges these strengths, combining CNN's feature extraction prowess with RNN's sequence processing capabilities.

3.6.2 Performance Metric

The authentication performance of the proposed PLA model is measured by authentication accuracy as:

$$\text{AucRate} = \frac{1}{N}\sum_{n=1}^{N}\mathbb{I}\left(\boldsymbol{L}_n = \boldsymbol{Y}_n\right) \quad (8),$$

where N is the number of CSI fingerprint sequences, and $\boldsymbol{L}_n$ and $\boldsymbol{Y}_n$ respectively stand for the real and predicted identity labels of the n-th CSI fingerprint sequence. If $\cdot$ is true, $\mathbb{I}(\cdot) = 1$; if $\cdot$ is false, $\mathbb{I}(\cdot) = 0$.

3.6.3 Simulation Parameters

CSI fingerprints are generated through the MATLAB platform, and the performance of the proposed scheme is verified through Python. The positions of users, RISs, and Bob are provided in Fig. 7, and the detailed parameters are provided in Table 2. The number of layers in GNNs typically depends on the complexity of the dataset. For a straightforward graph, just a few layers may suffice to capture valuable information. However, for intricate graph structures, more layers may be required to extract sophisticated feature representations. Furthermore, while increasing the number of layers can enhance the model's expressive power, it may also introduce issues such as over-fitting, where node characteristics converge and become indistinguishable after multiple layers of propagation, thereby impeding the model's ability to differentiate between nodes. Additionally, it may lead to problems like gradient vanishing or exploding. Consequently, in our simulation, the number of GNN layers is set to 3. The selection of the batch size should consider hardware resources, dataset size, and model complexity. Therefore, we choose a batch size of 16.

3.6.4 Simulation Results

Fig. 8 analyzes the authentication accuracy versus different distances between adjacent users. As the distance between users decreases, the similarity of CSI fingerprints increases,

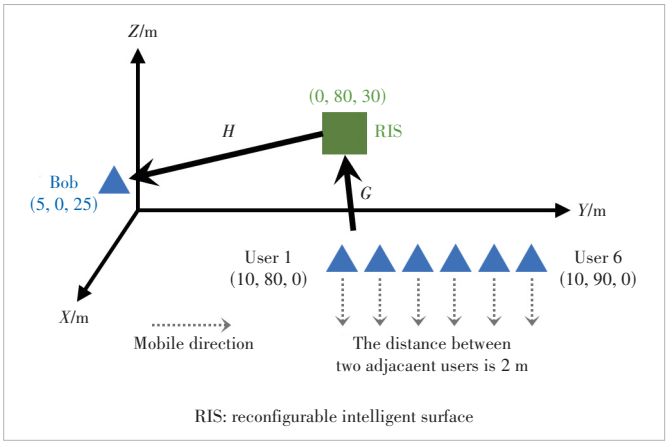


Figure 7. Positions of users, RISs, and Bob

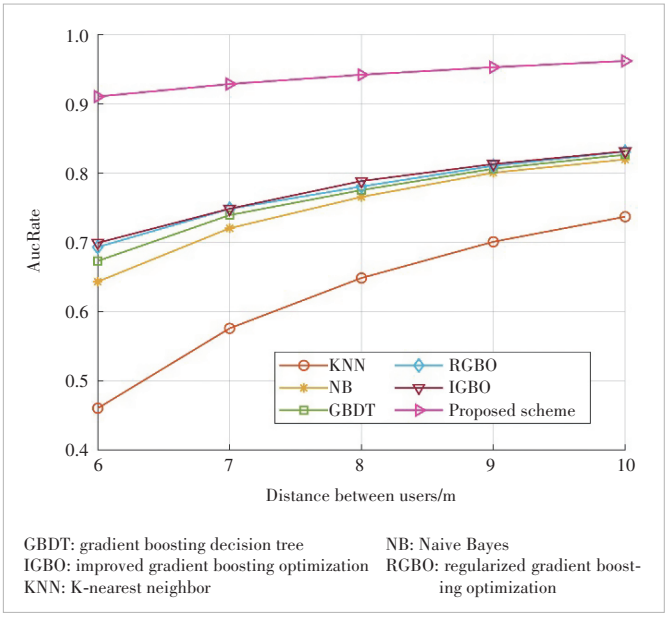
Table 2. Simulation parameters

Parameters	Values	Parameters	Values
N_T	4	N_T	3
Number of RIS elements	8×16	Carrier frequency	3.5 GHz
κ_H	3	κ_G	4
Bandwidth	1 MHz	Speed of users	2 m/s
Number of each user's CSI fingerprint samples	50 000	Number of each user's CSI fingerprint sequences	1 000
Length of each CSI fingerprint sequence	50	Ratio of training fingerprints	0.6
Learning rate	0.000 1	Batch size	16
Number of GNN layers	3	Ratio of pooling for nodes	0.2

CSI: channel state information GNN: graph neural network RIS: reconfigurable intelligent surface

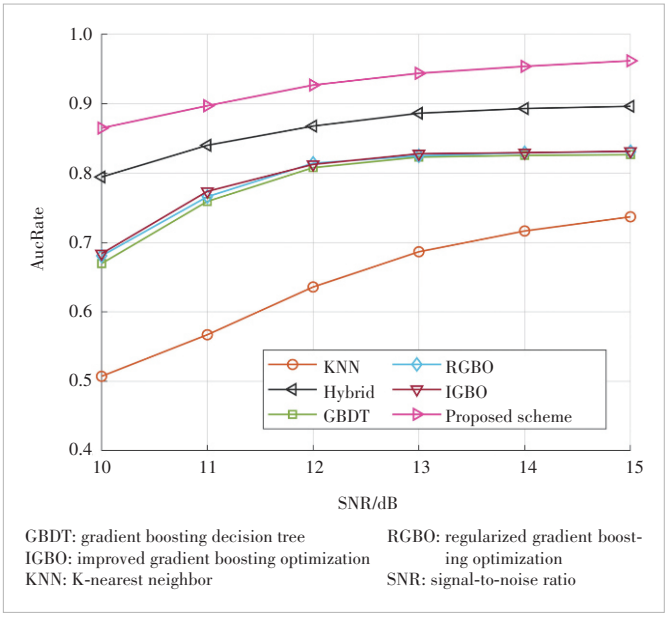
leading to a higher degree of overlap in their fingerprint distributions. Consequently, it becomes more challenging for the PLA model to distinguish between them, resulting in lower authentication accuracy. However, the proposed PLA scheme consistently outperforms the benchmark models.

Fig. 9 depicts the authentication accuracy versus different SNRs. The authentication accuracy of baseline schemes improves gradually with higher SNRs. Regardless of SNR levels, the proposed scheme consistently outperforms these baselines, demonstrating superior robustness. This superiority stems from its consideration of the variations in CSI fingerprint distribution caused by user movements, whereas the other methods presume an independent and identical distribution of CSI fingerprints for each user.



GBDT: gradient boosting decision tree
IGBO: improved gradient boosting optimization
KNN: K-nearest neighbor
NB: Naive Bayes
RGBO: regularized gradient boosting optimization

Figure 8. Authentication accuracy versus different distances between adjacent users



GBDT: gradient boosting decision tree
IGBO: improved gradient boosting optimization
KNN: K-nearest neighbor
RGBO: regularized gradient boosting optimization
SNR: signal-to-noise ratio

Figure 9. Authentication accuracy versus different SNRs

4 Future Research Directions

This section gives challenges and the future research direction of AI-driven PLA, including semantic fingerprint-based PLA, large AI model-based PLA, cross-layer PLA, multimodal signature-based PLA, distributed autonomous PLA, and PLA for emerging applications.

4.1 Semantic Fingerprint-Based PLA

Unlike traditional syntax-based communication paradigms that focus on indiscriminate transmission of bit data, semantic communications ensure an accurate understanding of the communication intent of source information at both the transmitting and receiving ends through the representation and measurement of semantic information, on-demand compression, and efficient and robust transmission. Inspired by semantic communications, we can extract knowledge of environmental semantic features from the channel propagation environment. By doing so, the physical channel can be abstracted as a semantic channel to assist in guiding the acquisition and optimization of channel fingerprints. Ref. [42] proposes an environmental semantics-enabled PLA method, which extracts frequency-independent wireless channel fingerprints from CSI in massive MIMO systems based on environmental semantic knowledge. The proposed method can effectively detect physical-layer spoofing attacks and is robust in time-varying wireless environments. In the future, constructing a knowledge base of semantic channel fingerprints and a semantic channel knowledge map can further enhance the efficiency and accuracy of PLA.

4.2 Large AI Model-Based PLA

In recent years, research on large models has been in full swing, and they offer the following advantages. 1) Large models possess more parameters, enabling them to learn more complex data patterns and thus perform better on various tasks. 2) The knowledge learned by large models during training is more generalizable, allowing for better generalization to unseen data and reducing the need for extensive labeled data. 3) With ongoing advancements in computing resources, the cost of training and deploying large models has gradually decreased. In the future, for multiuser authentication needs, high-robustness authentication requirements in complex environments, and lightweight authentication needs, PLA empowered by large models will exhibit exceptional performance.

4.3 Cross-Layer PLA

The training of PLA models based on AI requires the guidance of prior knowledge of legitimate fingerprints, which originates from identity labeling by upper-layer authentication mechanisms. Therefore, PLA is a type of cross-layer authenti-

cation technology, and its complexity is influenced by the interaction efficiency between the upper layer and the physical layer. Ref. [43] deploys active learning to select optimal unlabeled fingerprints and queries the identity from the upper-layer authentication protocol. The proposed method can effectively reduce the interaction requirements between the upper-layer and the physical-layer, achieving efficient utilization of prior fingerprint information. In the future, optimizing the fingerprint selection algorithm could further reduce the authentication error rate while maintaining lightweight performance.

4.4 Multimodal Signature-Based PLA

By integrating technologies such as wireless communications, radio sensing, and even AI, integrated sensing and communication (ISAC) can achieve the goals of spectrum conservation, cost reduction, and mutual enhancement between communication and sensing. Ref. [44] introduces the concept of synesthesia of machines and establishes a platform for generating and collecting communication and multimodal sensing information. This platform can provide multimodal data under diverse scenarios (urban, suburban, and rural) and various conditions (different weather, times of day, traffic densities, frequency bands, and antenna arrays). In the future, by designing multimodal fusion algorithms that integrate channel fingerprints, RF sensing data (millimeter-wave radar point clouds), and non-RF sensing data (RGB images, depth maps, and LiDAR point clouds), highly reliable identity authentication in dynamic and complex environments can be achieved.

4.5 Distributed Autonomous PLA

With the advancement in cloud computing and edge intelligence, the cloud-edge-end collaborative architecture can optimize resource utilization in a distributed manner and enhance data security. Ref. [45] proposes a privacy-preserving collaborative authentication scheme that provides reliable and efficient security, improved robustness in dynamic or untrusted environments, and stronger defensive capabilities compared with traditional centralized authentication methods. Future research includes cross-domain distributed PLA systems to ensure seamless switching and access for users or devices across different domains.

4.6 PLA for Emerging Applications

Future 6G networks will expand the boundaries of communication technology and transform the way we live and work. 6G will support emerging application scenarios, such as integrated space-air-ground-sea networks for ubiquitous coverage. Ref. [46] considers the identity security of satellite transmitters and provides a PLA scheme for low-earth orbit satellites. Ref. [47] provides a PLA approach for complicated time-varying underwater acoustic channels. Future research includes optimizing fingerprint feature extraction algorithms, developing anti-interference PLA technologies, and assessing industrial feasibility.

5 Conclusions

As the next generation of mobile communication technology, 6G stands as a pinnacle of global technological advancement and plays a pivotal role in driving future industrial development. As the latest iteration of information infrastructure, the security of 6G directly relates to the safe operation of national critical infrastructures. Currently, authentication mechanisms in wireless communications primarily rely on cryptography-based algorithms, and these "add-on" and "patchwork" authentication mechanisms face challenges in terms of security protection levels, computational power requirements, and compatibility. As an endogenous security approach, AI-based PLA boasts strong security assurance, intelligence, efficiency, and strong scalability. This paper first reviews representative AI-enabled PLA schemes, categorizing them into RF fingerprint extraction, fingerprint data augmentation, lightweight authentication models, authentication parameter optimization, multi-attacker identification, and physical-layer key generation for FDD systems. Furthermore, this paper proposes a GNN-based solution to identifing mobile multiusers and compares its performance with six baseline schemes to verify its superiority. Finally, this paper outlines future research directions, providing new insights for researchers in related fields.

References

[1] CHAFII M, BARIAH L, MUHAIDAT S, et al. Twelve scientific challenges for 6G: rethinking the foundations of communications theory [J]. IEEE communications surveys and tutorials, 2023, 25(2): 868 – 904. DOI: 10.1109/COMST.2023.3243918
[2] NGUYEN V L, LIN P C, CHENG B C, et al. Security and privacy for 6G: a survey on prospective technologies and challenges [J]. IEEE communications surveys and tutorials, 2021, 23(4): 2384 – 2428. DOI: 10.1109/COMST.2021.3108618
[3] GUO H Z, LI J Y, LIU J J, et al. A survey on space-air-ground-sea integrated network security in 6G [J]. IEEE communications surveys and tutorials, 2022, 24(1): 53 – 87. DOI: 10.1109/COMST.2021.3131332
[4] WANG C X, YOU X H, GAO X Q, et al. On the road to 6G: visions, requirements, key technologies, and testbeds [J]. IEEE communications surveys and tutorials, 2023, 25(2): 905 – 974. DOI: 10.1109/COMST.2023.3249835
[5] PORAMBAGE P, GÜR G, OSORIO D P M, et al. The roadmap to 6G security and privacy [J]. IEEE open journal of the communications society, 2021, 2: 1094 – 1122. DOI: 10.1109/OJCOMS.2021.3078081
[6] CHORTI A, BARRETO A N, KÖPSELL S, et al. Context-aware security for 6G wireless: the role of physical layer security [J]. IEEE communications standards magazine, 2022, 6(1): 102 – 108. DOI: 10.1109/MCOMSTD.0001.2000082
[7] LI D M, YANG X, ZHOU F H, et al. Blind physical-layer authentication based on composite radio sample characteristics [J]. IEEE transactions on communications, 2022, 70(10): 6790 – 6803. DOI: 10.1109/TCOMM.2022.3200599
[8] WANG X B, HAO P, HANZO L. Physical-layer authentication for wireless security enhancement: current challenges and future developments [J]. IEEE communications magazine, 2016, 54(6): 152 – 158. DOI: 10.1109/MCOM.2016.7498103

[9] XIE N, TAN H J, HUANG L, et al. Physical-layer authentication in wirelessly powered communication networks [J]. IEEE/ACM transactions on networking, 2021, 29(4): 1827 – 1840. DOI: 10.1109/TNET.2021.3071670
[10] HAN S F, XIE T, I C L. Greener physical layer technologies for 6G mobile communications [J]. IEEE communications magazine, 2021, 59(4): 68 – 74. DOI: 10.1109/MCOM.001.2000484
[11] FANG H, WANG X B, TOMASIN S. Machine learning for intelligent authentication in 5G and beyond wireless networks [J]. IEEE wireless communications, 2019, 26(5): 55 – 61. DOI: 10.1109/MWC.001.1900054
[12] FANG H, QI A, WANG X B. Fast authentication and progressive authorization in large-scale IoT: how to leverage AI for security enhancement [J]. IEEE network, 2020, 34(3): 24 – 29. DOI: 10.1109/MNET.011.1900276
[13] XIA S D, TAO X F, LI N, et al. Multiple correlated attributes based physical layer authentication in wireless networks [J]. IEEE transactions on vehicular technology, 2021, 70(2): 1673 – 1687. DOI: 10.1109/TVT.2021.3055563
[14] JIAN T, RENDON B C, OJUBA E, et al. Deep learning for RF fingerprinting: a massive experimental study [J]. IEEE Internet of Things magazine, 2020, 3(1): 50 – 57. DOI: 10.1109/IOTM.0001.1900065
[15] MENG R, XU B X, XU X D, et al. A survey of machine learning-based physical-layer authentication in wireless communications [J]. Journal of network and computer applications, 2025, 235: 104085. DOI: 10.1016/j.jnca.2024.104085
[16] PENG L N, ZHANG J Q, LIU M, et al. Deep learning based RF fingerprint identification using differential constellation trace figure [J]. IEEE transactions on vehicular technology, 2020, 69(1): 1091 – 1095. DOI: 10.1109/TVT.2019.2950670
[17] ROY D, MUKHERJEE T, CHATTERJEE M, et al. RF transmitter fingerprinting exploiting spatio-temporal properties in raw signal data [C]//Proceedings of 18th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2019: 89 – 96. DOI: 10.1109/icmla.2019.00023
[18] ZENG Y, GONG Y, LIU J W, et al. Multi-channel attentive feature fusion for radio frequency fingerprinting [J]. IEEE transactions on wireless communications, 2024, 23(5): 4243 – 4254. DOI: 10.1109/TWC.2023.3316286
[19] GOPALAKRISHNAN S, CEKIC M, MADHOW U. Robust wireless fingerprinting via complex-valued neural networks [C]//Proceedings of IEEE Global Communications Conference (GLOBECOM). IEEE, 2019: 1 – 6. DOI: 10.1109/globecom38437.2019.9013154
[20] MENG R, XU X D, SUN H, et al. Multiuser physical-layer authentication based on latent perturbed neural networks for industrial Internet of Things [J]. IEEE Internet of Things journal, 2023, 10(1): 637 – 652. DOI: 10.1109/JIOT.2022.3203514
[21] LIAO R F, WEN H, CHEN S L, et al. Multiuser physical layer authentication in Internet of Things with data augmentation [J]. IEEE Internet of Things journal, 2020, 7(3): 2077 – 2088. DOI: 10.1109/JIOT.2019.2960099
[22] CHEN Y, HO P H, WEN H, et al. On physical-layer authentication via online transfer learning [J]. IEEE Internet of Things journal, 2022, 9(2): 1374 – 1385. DOI: 10.1109/JIOT.2021.3086581
[23] WANG Y, GUI G, GACANIN H, et al. An efficient specific emitter identification method based on complex-valued neural networks and network compression [J]. IEEE journal on selected areas in communications, 2021, 39(8): 2305 – 2317. DOI: 10.1109/JSAC.2021.3087243
[24] XIAO L, LI Y, HAN G A, et al. PHY-layer spoofing detection with reinforcement learning in wireless networks [J]. IEEE transactions on vehicular technology, 2016, 65(12): 10037 – 10047. DOI: 10.1109/TVT.2016.2524258
[25] XIAO L, LU X Z, XU T W, et al. Reinforcement learning-based physical-layer authentication for controller area networks [J]. IEEE transactions on information forensics and security, 2021, 16: 2535 – 2547. DOI: 10.1109/TIFS.2021.3056206
[26] SENIGAGLIESI L, BALDI M, GAMBI E. Comparison of statistical and machine learning techniques for physical layer authentication [J]. IEEE transactions on information forensics and security, 2020, 16: 1506 – 1521. DOI: 10.1109/TIFS.2020.3033454
[27] QIU X Y, JIANG T, WU S, et al. Physical layer authentication enhancement using a Gaussian mixture model [J]. IEEE access, 2018, 6: 53583 – 53592. DOI: 10.1109/ACCESS.2018.2871514
[28] ZHANG X W, LI G Y, ZHANG J Q, et al. Deep-learning-based physical-layer secret key generation for FDD systems [J]. IEEE Internet of Things journal, 2022, 9(8): 6081 – 6094. DOI: 10.1109/JIOT.2021.3109272
[29] JIN L, XU X D, HAN S J, et al. RIS-assisted physical layer key generation and transmit power minimization [C]//Proceedings of IEEE Wireless Communications and Networking Conference (WCNC). IEEE, 2022: 2065 – 2070. DOI: 10.1109/WCNC51071.2022.9771815
[30] MENG R, XU X D, ZHAO H Y, et al. Multi-observation multi-channel-attribute-based multi-user authentication for industrial wireless edge networks [J]. IEEE transactions on industrial informatics, 2024, 20(2): 2097 – 2108. DOI: 10.1109/TII.2023.3286885
[31] MOURYA S, REDDY P, AMURU S, et al. Spectral temporal graph neural network for massive MIMO CSI prediction [J]. IEEE wireless communications letters, 2024, 13(5): 1399 – 1403. DOI: 10.1109/LWC.2024.3372148
[32] ZHENG B X, YOU C S, MEI W D, et al. A survey on channel estimation and practical passive beamforming design for intelligent reflecting surface aided wireless communications [J]. IEEE communications surveys & tutorials, 2022, 24(2): 1035 – 1071. DOI: 10.1109/COMST.2022.3155305
[33] RODRÍGUEZ P, BAUTISTA M A, GONZÀLEZ J, et al. Beyond one-hot encoding: lower dimensional target embedding [J]. Image and vision computing, 2018, 75: 21 – 31. DOI: 10.1016/j.imavis.2018.04.004
[34] WU Z H, PAN S R, CHEN F W, et al. A comprehensive survey on graph neural networks [J]. IEEE transactions on neural networks and learning systems, 2021, 32(1): 4 – 24. DOI: 10.1109/TNNLS.2020.2978386
[35] ZHOU J, CUI G Q, HU S D, et al. Graph neural networks: a review of methods and applications [J]. AI open, 2020, 1: 57 – 81. DOI: 10.1016/j.aiopen.2021.01.001
[36] LEA C, FLYNN M D, VIDAL R, et al. Temporal convolutional networks for action segmentation and detection [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 1003 – 1012. DOI: 10.1109/CVPR.2017.113
[37] SKARDING J, GABRYS B, MUSIAL K. Foundations and modeling of dynamic networks using dynamic graph neural networks: a survey [J]. IEEE access, 2021, 9: 79143 – 79168. DOI: 10.1109/ACCESS.2021.3082932
[38] LIU H Y, YANG D H, LIU X Z, et al. TodyNet: temporal dynamic graph neural network for multivariate time series classification [J]. Information sciences, 2024, 677: 120914. DOI: 10.1016/j.ins.2024.120914
[39] PAN F, PANG Z B, WEN H, et al. Threshold-free physical layer authentication based on machine learning for industrial wireless CPS [J]. IEEE transactions on industrial informatics, 2019, 15(12): 6481 – 6491. DOI: 10.1109/TII.2019.2925418
[40] WEBB G I. Naive bayes [M]//Encyclopedia of machine learning. Boston, USA: Springer, 2011: 713 – 714. DOI: 10.1007/978-0-387-30164-8_576
[41] ALZAHRANI S, ALDERAAN J, ALATAWI D, et al. Continuous mobile user authentication using a hybrid CNN-Bi-LSTM approach [J]. Computers, materials & continua, 2023, 75(1): 651 – 667. DOI: 10.32604/cmc.2023.035173
[42] GAO N, HUANG Q Y, LI C, et al. EsaNet: environment semantics enabled physical layer authentication [J]. IEEE wireless communications letters, 2024, 13(1): 178 – 182. DOI: 10.1109/LWC.2023.3324981
[43] MENG R, ZHU F Z, XU X D, et al. Efficient Gaussian process classification-based physical-layer authentication with configurable fingerprints for 6G-enabled IoT [EB/OL]. [2024-11-10]. https://arxiv.org/abs/2307.12263v2
[44] CHENG X, HUANG Z W, BAI L, et al. M^3SC: a generic dataset for mixed multi-modal (MMM) sensing and communication integration [J]. China communications, 2023, 20(11): 13 – 29. DOI: 10.23919/JCC.fa.2023-

0268.202311
[45] FANG H, WANG X B, XIAO Z L, et al. Autonomous collaborative authentication with privacy preservation in 6G: from homogeneity to heterogeneity [J]. IEEE network, 2022, 36(6): 28 – 36. DOI: 10.1109/MNET.002.2100312
[46] OLIGERI G, SCIANCALEPORE S, RAPONI S, et al. PAST-AI: physical-layer authentication of satellite transmitters via deep learning [J]. IEEE transactions on information forensics and security, 2022, 18: 274 – 289. DOI: 10.1109/TIFS.2022.3219287
[47] ZHAO R Q, SHI T, LIU C Y, et al. Physical layer authentication without adversary training data in resource-constrained underwater acoustic networks [J]. IEEE sensors journal, 2023, 23(22): 28270 – 28281. DOI: 10.1109/JSEN.2023.3321777

Biographies

MENG Rui received his BS degree in information engineering and PhD degree in information and communication engineering both from Beijing University of Posts and Telecommunications (BUPT), China in 2018 and 2024, respectively. He is currently a postdoctoral fellow with BUPT. His research interests cover next-generation networks, physical layer authentication, identity security, semantic security, deep learning, and Internet of Things.

FAN Dayu received his BS degree in information engineering from Beijing University of Posts and Telecommunications (BUPT), China in 2024, where he is currently pursuing his master's degree in communication engineering. His research interests cover wireless security, semantic communication, and deep learning.

XU Xiaodong (xuxiaodong@bupt.edu.cn) received his BS degree in information and communication engineering and master's degree in communication and information system both from Shandong University, China in 2001 and 2004, respectively. He received his PhD degree in circuit and system from Beijing University of Posts and Telecommunications (BUPT), China in 2007. He is currently a professor of BUPT, a research fellow of the Department of Broadband Communication of Peng Cheng Laboratory and a member of IMT-2030 (6G) Experts Panel. He has coauthored nine books/chapters and more than 120 journal and conference papers. He is also the inventor or co-inventor of 51 granted patents. His research interests cover semantic communications, intellicise communication systems, moving networks, and mobile edge computing and caching.

LYU Suyu received her bachelor's degree and PhD degree in information and communication engineering from Beijing University of Posts and Telecommunications, China in 2018 and 2024, respectively. From November 2022 to September 2023, she was a visiting student with the School of Electronic Engineering and Computer Science, Queen Mary University of London, UK. She is currently a post-doctoral researcher at Beijing University of Technology, China. Her main research interests include ultra-reliable low-latency communications, reconfigurable intelligent surface, and non-orthogonal multiple access.

TAO Xiaofeng received his BS degree in electrical engineering from Xi'an Jiaotong University, China in 1993, and MS and PhD degrees in telecommunication engineering from Beijing University of Posts and Telecommunications (BUPT), China, in 1999 and 2002, respectively. He is a professor at BUPT, a fellow of the IET, and Chair of the IEEE ComSoc Beijing Chapter. He has authored or co-authored over 200 papers and three books in wireless communication areas. He focuses on 5G/B5G research.

Separate Source Channel Coding Is Still What You Need: An LLM-Based Rethinking

REN Tianqi[1], LI Rongpeng[1], ZHAO Mingmin[1], CHEN Xianfu[2], LIU Guangyi[3], YANG Yang[4], ZHAO Zhifeng[1,5], ZHANG Honggang[6]

(1. College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China;
2. Shenzhen CyberAray Network Technology Co., Ltd., Shenzhen 518000, China;
3. China Mobile Research Institute, Beijing 100053, China;
4. The Internet of Things Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511453, China;
5. Zhejiang Lab, Hangzhou 311121, China;
6. Faculty of Data Science, City University of Macau, Macao 999078, China)

DOI: 10.12142/ZTECOM.202501005

https://kns.cnki.net/kcms/detail/34.1294.TN.20250319.0931.004.html, published online March 20, 2025

Manuscript received: 2025-01-02

Abstract: Along with the proliferating research interest in semantic communication (SemCom), joint source channel coding (JSCC) has dominated the attention due to the widely assumed existence in efficiently delivering information semantics. Nevertheless, this paper challenges the conventional JSCC paradigm and advocates for adopting separate source channel coding (SSCC) to enjoy a more underlying degree of freedom for optimization. We demonstrate that SSCC, after leveraging the strengths of the Large Language Model (LLM) for source coding and Error Correction Code Transformer (ECCT) complemented for channel coding, offers superior performance over JSCC. Our proposed framework also effectively highlights the compatibility challenges between SemCom approaches and digital communication systems, particularly concerning the resource costs associated with the transmission of high-precision floating point numbers. Through comprehensive evaluations, we establish that assisted by LLM-based compression and ECCT-enhanced error correction, SSCC remains a viable and effective solution for modern communication systems. In other words, separate source channel coding is still what we need.

Keywords: separate source channel coding (SSCC); joint source channel coding (JSCC); end-to-end communication system; Large Language Model (LLM); lossless text compression; Error Correction Code Transformer (ECCT)

Citation (Format 1): REN T Q, LI R P, ZHAO M M, et al. Separate source channel coding is still what you need: an LLM-based rethinking [J]. ZTE Communications, 2025, 23(1): 30 - 44. DOI: 10.12142/ZTECOM.202501005

Citation (Format 2): T. Q. Ren, R. P. Li, M. M. Zhao, et al., "Separate source channel coding is still what you need: an LLM-based rethinking," *ZTE Communications*, vol. 23, no. 1, pp. 30 - 44, Mar. 2025. doi: 10.12142/ZTECOM.202501005.

1 Introduction

Semantic communication (SemCom) has garnered significant attention in recent years, with researchers exploring innovative approaches to enhance the efficiency and reliability of information transmission[1]. Generally, SemCom leverages deep learning-based joint source-channel coding (JSCC) methods to preserve global semantic information and local texture during the transmission process. DeepJSCC[2] pioneers these works by implementing JSCC with feedback and allowing for real-time adaptation to channel conditions. Along with its steady progress, JSCC has been substantially studied, mostly with the optimization objective shifting from bit error rates to the semantic relevance of the transmitted information in SemCom[3-14]. However, albeit the awfully exploded research interest, one critical question remains unsolved: why does the joint approach stand out, as separate source channel coding (SSCC), shall promise a greater degree of freedom from an optimization perspective?

As the terminology implies, SSCC encompasses two decoupled ingredients: source coding and channel coding. The former part lies in effectively compressing the context, and the effectiveness of underlying deep neural networks (DNN)-based predictors, such as recurrent neural networks (RNN)-

This work was supported in part by the National Key Research and Development Program of China under Grant No. 2024YFE0200600, the Zhejiang Provincial Natural Science Foundation of China under Grant No. LR23F010005, and the Huawei Cooperation Project under Grant No. TC20240829036.

based DeepZip[15], Long Short Term Memory (LSTM) - based[16 - 17] and hybrid DNN-based Dzip[18], have been validated widely in achieving satisfactory text compression. More prominently, Transformer-based[19] and Large Language Model (LLM)-based compression have emerged recently[20 - 24]. The latest research[25] unveils the equivalence between compression and prediction. In other words, in the general framework where statistical models predict symbols and encoders use predictive probabilities to perform compression, better predictive models lead directly to better compressors[25]. Hence, the astonishing capability of LLM implies the potential for an unprecedented source codec. On the other hand, the Error Correction Code (ECC) plays an indispensable role in channel coding. Although some advanced algebraic block codes like Bose - Chaudhuri - Hocquenghem (BCH) codes[26], Low-Density Parity-Check (LDPC) codes[27] and Polar codes[28] can somewhat ensure the reliability of transmission, the efficient decoding of ECC is an unresolved difficulty. Recently, DNNs have started to demonstrate their contribution to channel coding. For example, deep learning models are implemented to achieve belief propagation (BP) decoding[29 - 31], while a model-free Error Correction Code Transformer (ECCT) for algebraic block codes[32] contributes to the enhancement of decoding reliability.

In this paper, on top of an LLM-based arithmetic coding (LLM-AC) system, the proposed SSCC framework integrates fine-grained, semantics-aware probability modeling and encoding with ECCT-enhanced channel decoding, thus forming a closed-loop optimization framework. To the best of our knowledge, this work represents the first comprehensive integration of LLM-based compression and ECCT-complemented channel decoding for a holistic SemCom architecture. Through extensively showcasing the performance superiority over JSCC, we argue this performance improvement primarily arises after tackling the underlying incompatibility between conventional SemCom approaches[3 - 7, 11 - 14] and digital communication architectures[33]. Particularly, those approaches simply assume the deliverability of encoded semantic feature vectors while neglecting the energy costs associated with transmitting high-precision floating point numbers[33]. However, further quantization[9 - 10] and digital modulation can compromise the widely assumed existence of performance superiority in JSCC. Meanwhile, in contrast to the direct utilization of the astonishing semantic interpretation capability[34 - 36], the deployment of LLMs focuses on the compression and encoding of text to squeeze the largely untapped redundancy. Therefore, our work is also significantly different from existing integrations of generative AI (GAI) and SemCom[37 - 43]. Furthermore, the adoption of ECCT boosts the effectiveness of SSCC in specific cases. In summary, our comprehensive evaluation of LLM and ECCT-based SSCC demonstrates that separate source channel coding is still what we need.

The rest of this paper is organized as follows. Section 2 introduces the SSCC system model, while its key components are enumerated in Section 3. Section 4 provides numerical results demonstrating the performance superiority of the proposed SSCC system. Finally, Section 5 concludes this paper with discussions on future works. For convenience, we list the major notations of this paper in Table 1.

2 System Model

Our SSCC framework encompasses the following ingredients.

1) Source encoding

The input text sequence denoted as $\boldsymbol{s}$ undergoes a source encoder that converts characters into a compressed binary message $\boldsymbol{m} \in \{0,1\}^K$. During source encoding, arithmetic coding (AC) can be leveraged for effective compression here. For LLM-based processing, an intermediate result (i.e., a sequence of tokens $\boldsymbol{t}$) can be obtained during the transformation from $\boldsymbol{s}$ to $\boldsymbol{m}$.

Table 1. Major notations used in this paper

Notation	Definition
$\boldsymbol{s}$, $\hat{\boldsymbol{s}}$	The transmitted text sequence and the recovered text sequenceat the receiver side
$\boldsymbol{t}$, $\hat{\boldsymbol{t}}$	The transmitted token sequence and the recovered token sequence at the receiver side
C_s, C_e	The source code and the channel code (error correction code)
ρ, $\tilde{\rho}$	The source distribution and the predicted probability distribution via LLM
$\mathcal{D}$, D_i, τ	The dictionary of source coder, the i-th character in the dictionary, and the vocabulary of the dictionary
$\mathbb{I}_k$, l_k, u_k	The probability interval in step k of source coding and its corresponding lower and upper bounds
$\boldsymbol{m}$, $\hat{\boldsymbol{m}}$	The message encoded by the source coder and the received (and channel decoded) message
λ	The probability interval, determined by the codeword, in a decimal form
N, K	The codeword length and message length of error correction code $C_e(N,K)$
$\boldsymbol{G}$, $\boldsymbol{H}$	The generator matrix and the parity check matrix
$\boldsymbol{x}$, $\boldsymbol{x}_b$, $\boldsymbol{x}_s$	The transmitted codeword encoded by the channel coder and its binary and sign form
$\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{x}}_b$	The soft approximation of codeword and its binary form
$\mathcal{N}(\cdot,\cdot)$, σ_n	The Gaussian distribution and the standard deviation of noise
h	The channel fading coefficient
z, $\tilde{z}$, $\hat{z}$	The additive Gaussian noise, as well as its corresponding multiplicative noise and the prediction result by ECCT
$\boldsymbol{y}$, $\boldsymbol{y}_b$, $\tilde{\boldsymbol{y}}$	The noisy codeword, its binary form, and the result of pre-processing noisy codeword
$\mathrm{syn}(\cdot)$	The syndrome of codes defined in ECCT
$f(\cdot)$	The decoding function of ECCT
$\boldsymbol{W}$	The learnable embedding matrix for high-dimensional mapping
$\boldsymbol{g}(\cdot)$	The code-aware self-attention mask

ECCT: Error Correction Code Transformer LLM: Large Language Model

2) Channel encoding and modulation

The message $\boldsymbol{m}$ is then encoded via an LDPC code $C_e(N,K)$, which is selected for its excellent error-correction capabilities and compatibility with iterative decoding algorithms, as mentioned in Ref. [32]. The encoding process employs a generator matrix $\boldsymbol{G}$ to transform the message in $\boldsymbol{m}$ to a codeword $\boldsymbol{x}_b \in \{0,1\}^N$. The parity check matrix $\boldsymbol{H}$, which satisfies $\boldsymbol{G} \cdot \boldsymbol{H}^T = 0$ and $\boldsymbol{H} \cdot \boldsymbol{x}_b = 0$, is a key component of the LDPC decoding process. Afterwards, binary phase shift keying (BPSK) modulation maps the binary codeword $\boldsymbol{x}_b$ to a sequence of symbols $\boldsymbol{x}_s \in \{\pm 1\}^N$, suitable for transmission over the wireless channel. Notably, other error correction codes, such as Polar codes[28], can be applied as well.

3) Channel

The modulated signal $\boldsymbol{x}_s$ is transmitted over a noisy channel, modeled as an additive white Gaussian noise (AWGN) channel or a Rayleigh fading channel. The received signal $\boldsymbol{y} \in \mathbb{R}^N$ is corrupted by additive noise $\boldsymbol{z} \sim \mathcal{N}(0, \sigma_n^2)$, resulting in $\boldsymbol{y} = h\boldsymbol{x}_s + \boldsymbol{z}$, where h is the channel fading coefficient.

4) Demodulation and channel decoding

BPSK demodulation recovers a binary codeword $\hat{\boldsymbol{x}}_b \in \{0,1\}^N$ from $\hat{\boldsymbol{x}}$. Subsequently, the channel decoder reconstructs the message $\hat{\boldsymbol{m}} \in \{0,1\}^K$ from $\hat{\boldsymbol{x}}_b$. In contrast to conventional approaches that employ either hard-decision (e.g., the bit-flipping algorithm) or soft-decision (e.g., the sum-product algorithm) algorithms to decode LDPC codewords transmitted through the channel, some complementary decoding modules, such as ECCT, can be applied prior to demodulation to enhance the decoding performance. Notably, ECCT can provide an estimation of the transmitted codeword $\hat{\boldsymbol{x}}_b$, denoted as $\hat{\boldsymbol{x}}$, while subsequent demodulation and information bits extraction are then performed on the estimated codeword $\hat{\boldsymbol{x}}$.

5) Source decoding

The recovered message $\hat{\boldsymbol{m}}$ is ultimately decoded by the source decoder, which reconstructs the text sequence $\hat{\boldsymbol{s}}$ from the message, effectively reversing the encoding process. Similar to the encoder, the decoder can implement arithmetic decoding.

In comparison, JSCC typically employs an end-to-end DNN to implement source and channel codecs. Here, the terminology "end-to-end" implies the joint training of source and channel codes, as adopted in most works. Further details on JSCC can be found in Ref. [1] and the references therein. In the following section, we will address how to leverage the strength of LLM to enhance text compression and reconstruction, combined with the robustness of ECCT-complemented LDPC codes for error correction, as shown in Fig. 1.

3 Proposed SSCC Framework

In this section, we introduce LLM-based source coding and ECCT-complemented channel coding.

3.1 LLM-Based Source Coding

Given a source distribution ρ, lossless compression aims to encode a text sequence $\boldsymbol{s}$ sampled from ρ into a binary code $\boldsymbol{m} = C_s(\boldsymbol{s})$ of minimal possible length with no loss of original infor-

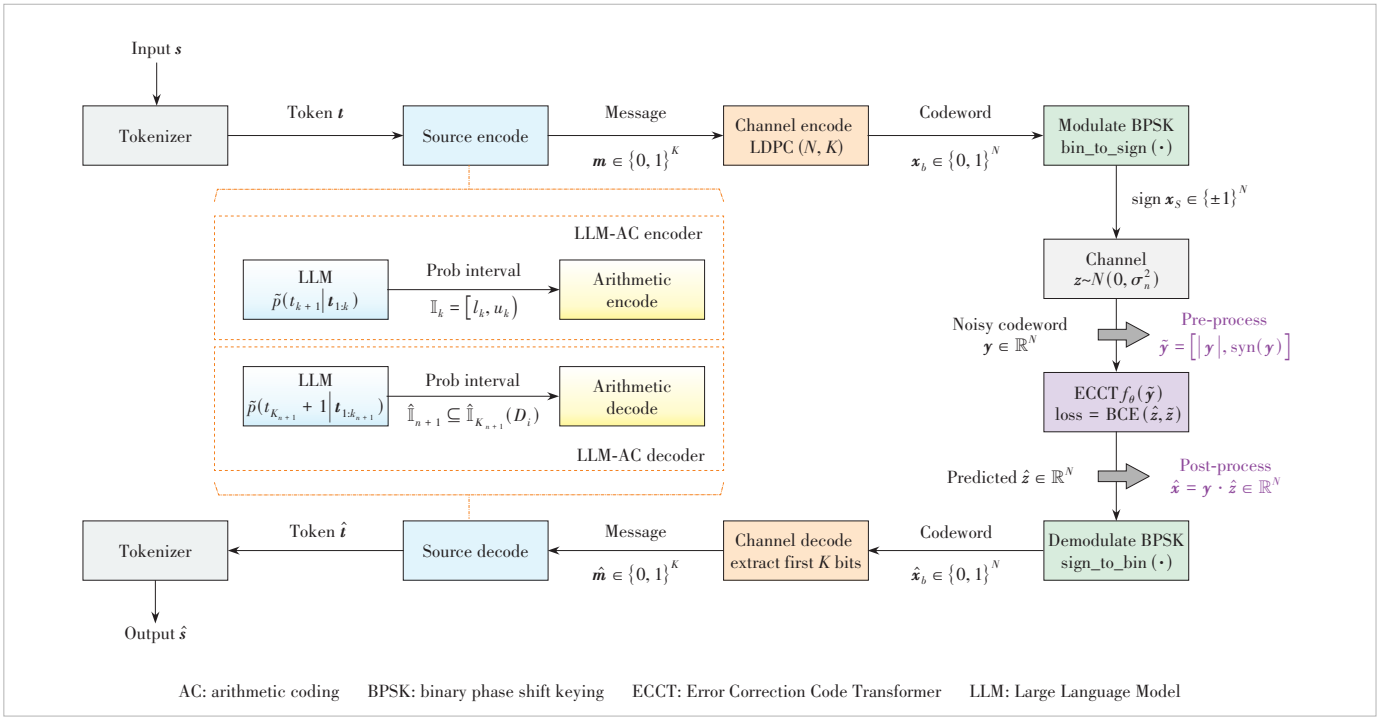


Figure 1. Framework of LLM-based and ECCT-complemented SSCC system

mation. According to Shannon's source coding theorem[44], the optimal expected bit length is $L_{\min} = \mathbb{E}_{\boldsymbol{s}\sim\rho}\left[-\log_2\rho(\boldsymbol{s})\right]$. To obtain such optimal length, arithmetic coding[45-46], a form of entropy encoding, is typically adopted, relying on a probabilistic model over ρ or its marginal distribution. Arithmetic coding implies that frequently used characters are stored with fewer bits while rarely occurring characters correspond to more bits, resulting in fewer bits used in total.

In particular, the input text sequence $\boldsymbol{s}$ undergoes tokenization by the LLM tokenizer, which converts characters into a sequence of tokens $\boldsymbol{t}$ for processing by the LLM. The LLM subsequently generates a compact representation of the text, effectively encoding the tokens into a compressed binary message $\boldsymbol{m} \in \{0,1\}^K$. Specially, considering a dictionary $\mathcal{D}$ of τ tokens, the input sequence $\boldsymbol{s}$ is first parsed into the token sequence $\boldsymbol{t}$. Given the first k tokens $\boldsymbol{t}_{1:k}$, the $(k+1)$-th token t_{k+1} can be inferred as a predicted probability distribution $\tilde{\rho}\left(t_{k+1}|\boldsymbol{t}_{1:k}\right)$. Here, $\tilde{\rho}\left(t_{k+1}|\boldsymbol{t}_{1:k}\right)$ indicates the LLM's estimation of the true distribution $\rho\left(t_{k+1}|\boldsymbol{t}_{1:k}\right)$. The incremental decoding nature in LLM enables it to accurately predict the probability distribution of the next token based on known ones, thereby providing a sub-optimal estimation of the true distribution[25]. As shown in Fig. 2, selecting the next character effectively narrows down the probabilistic interval where the sequence is located, which means the code $\boldsymbol{m}$ is determined once the interval is fixed. Starting with $\mathbb{I}_0 = [0,1)$, the previous interval determined by $\boldsymbol{t}_{1:k}$ in step k is defined as $\mathbb{I}_k = \left[l_k, u_k\right)$. Therefore, denoting $p\left(t_{k+1} = D_j\right) = \tilde{\rho}\left(t_{k+1} = D_j|\boldsymbol{t}_{1:k}\right)$,

$$\mathbb{I}_{k+1}\left(D_i\right) = \begin{bmatrix} l_k + \left(u_k - l_k\right) \times \sum_{j<i} p\left(t_{k+1} = D_j\right), \\ l_k + \left(u_k - l_k\right) \times \sum_{j\leqslant i} p\left(t_{k+1} = D_j\right) \end{bmatrix} \tag{1}.$$

In practice, we consider finite precision arithmetic encoders, referring to Ref. [47], with pseudo-code provided in Appendix 1. Consequently, we can obtain a binary code $\boldsymbol{m} = C_s(\boldsymbol{s})$ of the shortest length, completely corresponding to the probability interval determined by the sequence. At the receiver side, if the receiver shares a consistent source distribution $\tilde{\rho}$ with the sender, given the received (and channel-decoded) bit sequence $\hat{m}$ corresponding to $C_s(\boldsymbol{s})$, we can decode $t_{K_{n+1}} = D_i \in \mathcal{D}$ by identifying D_i, such that

$$\hat{\mathbb{I}}_{n+1} = \left[l_{n+1}, u_{n+1}\right) = \begin{cases} \left[l_n, \frac{1}{2}\left(l_n + u_n\right)\right), \text{if } m_{n+1} = 0 \\ \left[\frac{1}{2}\left(l_n + u_n\right), u_n\right), \text{if } m_{n+1} = 1 \end{cases} \subseteq \hat{\mathbb{I}}_{K_{n+1}}\left(D_i\right) = \left[L, U\right) \tag{2},$$

where $L = l_{K_{n+1}} + \left(u_{K_{n+1}} - l_{K_{n+1}}\right) \times \sum_{j<i} p\left(t_{K_{n+1}+1} = D_j\right)$ and $U = l_{K_{n+1}} + \left(u_{K_{n+1}} - l_{K_{n+1}}\right) \times \sum_{j\leqslant i} p\left(t_{K_{n+1}+1} = D_j\right)$. For more details, please refer to Appendix 1.

Fig. 3 illustrates such LLM-based arithmetic encoding and decoding, where the LLM provides a probability interval according to the text sequence $\boldsymbol{s}$. Unlike the online setting, which trains the model on the data to be compressed, this paper assumes the availability of a well-trained LLM and employs it to compress different datasets, following the offline setting used in Ref. [24].

Remark 1: Ref. [25] figures out that the expected code length achieved by leveraging LLM as a compressor could be represented as the cross-entropy, that is,

$$H\left(\rho, \tilde{\rho}\right) \coloneqq \mathbb{E}_{\boldsymbol{s}\sim\rho}\left[\sum_{i=1}^{n} -\log_2\tilde{\rho}\left(s_i|\boldsymbol{s}_{<i}\right)\right] \tag{3},$$

where ρ is the source distribution and $\tilde{\rho}$ is the estimation of ρ via a parametric probabilistic model. Hence, the compres-

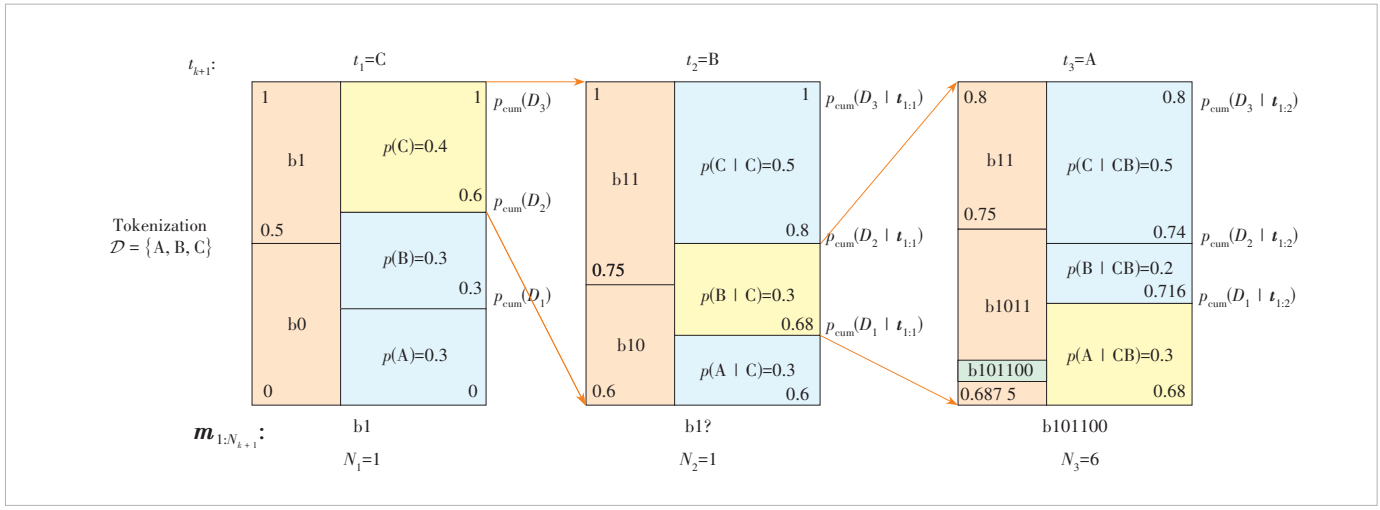


Figure 2. An example of arithmetic coding

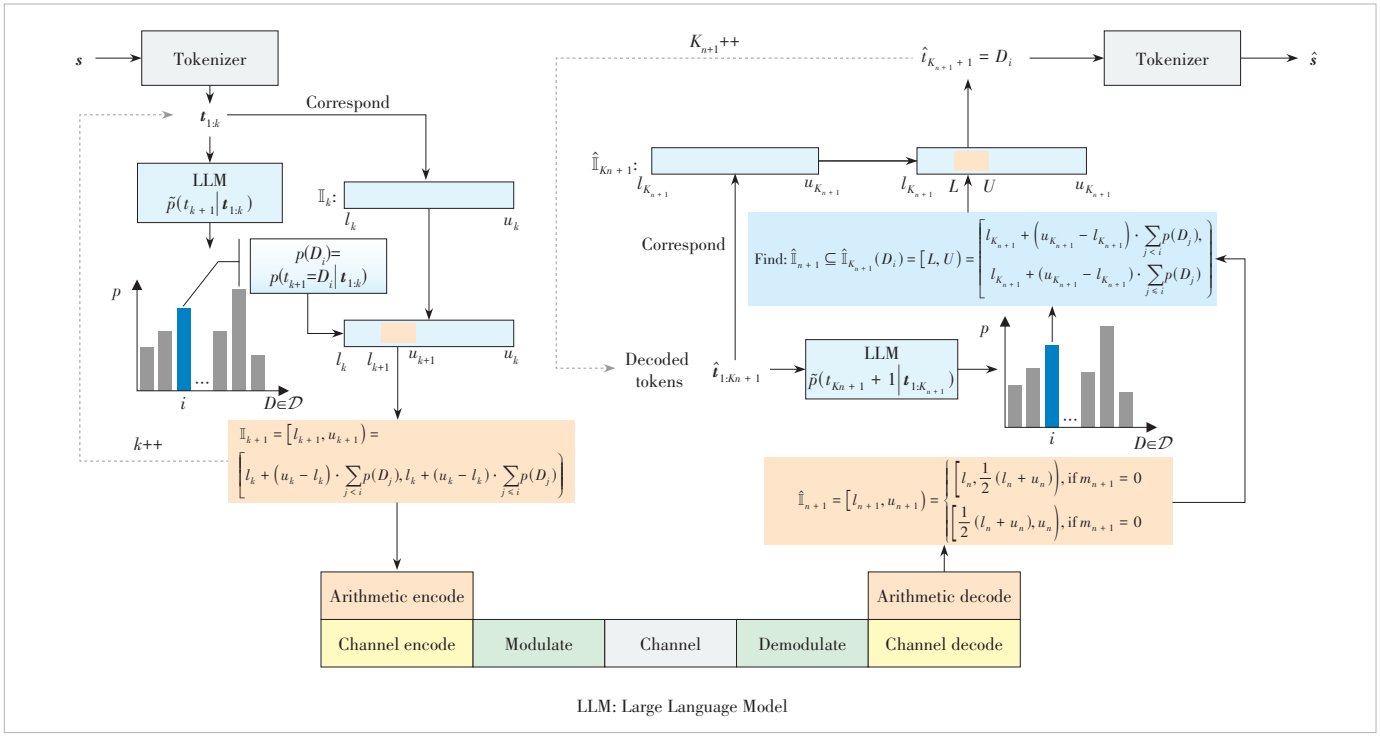


Figure 3. LLM-based arithmetic encoding and decoding

sion shares the same training objective as prediction. Therefore, it can be interpreted as the link between the model log-loss and the compression rate, providing theoretical support for the employment of LLM for source coding.

3.2 Error Correction Code Transformer

ECCT[48] belongs to the complementary Transformer-like module. It ensures the channel decoding reliability. Notably, ECCT involves specific preprocessing and post-processing steps to avoid overfitting effectively. Without the loss of generality, before preprocessing, the syndrome of codes is defined by

$$\mathrm{syn}(\boldsymbol{y}) \coloneqq \boldsymbol{H}\boldsymbol{y}_b = \boldsymbol{H}\mathrm{sign\_to\_bin}(\boldsymbol{y}) = \frac{1}{2}\boldsymbol{H}\left(1 - \mathrm{sign}(\boldsymbol{y})\right) \in \{0,1\}^{N-K} \tag{4}$$

This should be checked first upon receiving the signal since corruption could be detected immediately if $\mathrm{syn}(\boldsymbol{y})$ is a non-zero vector. In other words, an all-zero syndrome ensures that the received signal suffers no distortion. Note that the function $\mathrm{sign\_to\_bin}(\cdot)$ could be viewed as a hard decision on $\boldsymbol{y}$ and $\mathrm{sign}(\cdot)$ here denotes a sign function defined by

$$\mathrm{sign}(y)\begin{cases} 1, & y > 0 \\ 0, & y = 0 \\ -1, & y < 0 \end{cases} \tag{5}$$

Next, ECCT constructs a $2N - K$ dimensional input embedding by concatenating the element-wise magnitude and syndrome vectors, such that

$$\tilde{\boldsymbol{y}} \coloneqq \left[|\boldsymbol{y}|, \mathrm{syn}(\boldsymbol{y})\right] \in \mathbb{R}^{2N-K} \tag{6}$$

where $[\cdot,\cdot]$ denotes vector/matrix concatenation and $|\boldsymbol{y}|$ denotes the absolute value (magnitude) of $\boldsymbol{y}$.

The objective of the decoder is to predict the multiplicative noise $\tilde{\boldsymbol{z}}$ from $\boldsymbol{y}$, where $\boldsymbol{y} = h\boldsymbol{x}_s + \boldsymbol{z} = \boldsymbol{x}_s(h + \boldsymbol{x}_s\boldsymbol{z}) = \boldsymbol{x}_s\tilde{\boldsymbol{z}}$. Compared to traditional Transformer architectures[19], ECCT introduces two additional modules for positional reliability encoding and code aware self-attention, as shown in Fig. 4. Notably, ECCT processes the channel output $\boldsymbol{y}$ as input and generates a prediction $\hat{\boldsymbol{z}}$ of the multiplicative noise $\tilde{\boldsymbol{z}}$. The key differences between ECCT and traditional Transformer architectures are highlighted in the dashed-line boxes in Fig. 4. Implementation details are provided in Appendix 2.

Finally, the training process aims to minimize the binary cross entropy (BCE) loss between the predicted noise $\hat{\boldsymbol{z}}$ and the multiplicative noise $\tilde{\boldsymbol{z}}$, given by

$$\mathrm{loss} = \mathrm{BCELoss}(\hat{\boldsymbol{z}}, \tilde{\boldsymbol{z}}) = -\frac{1}{N}\sum_i \begin{pmatrix} \mathrm{bin}(\tilde{z}_i) \cdot \log\left(\sigma(\hat{z}_i)\right) + \\ \left(1 - \mathrm{bin}(\tilde{z}_i)\right) \cdot \log\left(1 - \sigma(\hat{z}_i)\right) \end{pmatrix} \tag{7}$$

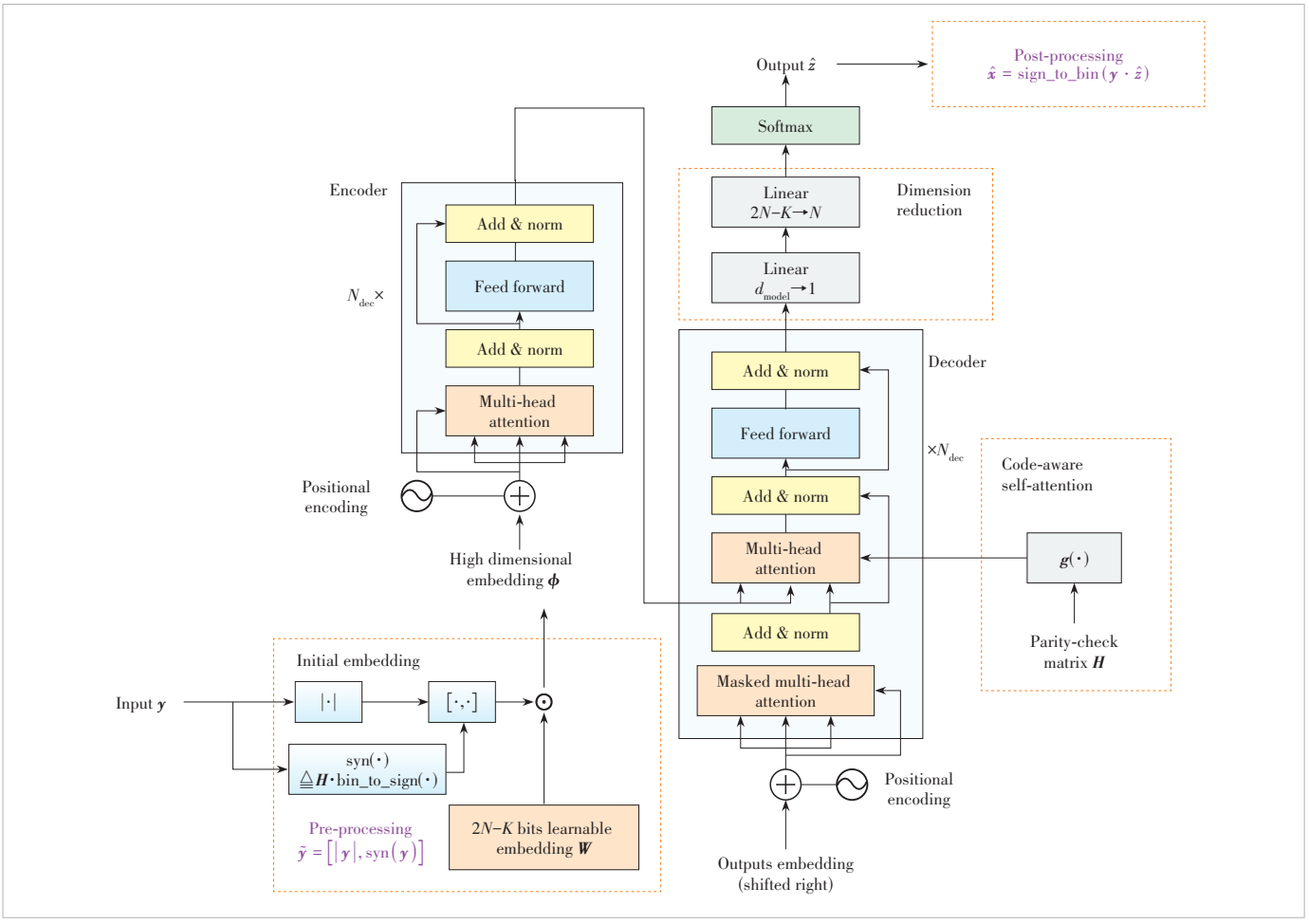


Figure 4. ECCT architecture

where $\sigma(\cdot)$ denotes the sigmoid activation function.

Remark 2: The estimation of multiplicative noise is represented as $\hat{z} = f(\tilde{y})$, while the post-processing step estimates x by $\hat{x} = \text{sign\_to\_bin}(y \cdot f(\tilde{y}))$. Given that, for correct estimation, $\text{sign}(\hat{z}) = \text{sign}(\tilde{z})$. Therefore,

$$\hat{x} = \text{sign\_to\_bin}(y \cdot f(\tilde{y})) = \text{sign\_to\_bin}(x_s \tilde{z} \cdot \hat{z}) = \text{sign\_to\_bin}(x_s) = x \quad (8).$$

In other words, ECCT contributes to noise-free channel coding.

4 Experiments

In this section, we compare the proposed method with traditional SSCC approaches and existing JSCC solutions under both AWGN and Rayleigh fading channels.

4.1 Simulation Settings

To facilitate comparison, we utilize a pre-processed dataset consisting of the standard proceedings of the European Parliament[49]. A segment of this dataset is selected as an example and fed as the source to a Generative Pre-Trained Transformer 2 (GPT2)[50] model for source coding. In this numerical experiment, we primarily choose the smallest GPT2-base model with 124 million parameters, while larger models (e. g., the 355-million-parameter GPT2-medium, the 774-million-parameter GPT2-large, and the1.5-billion-parameter GPT2-XL) are subsequently used for comparative analysis. Arithmetic coding based on the LLM is configured with a precision limit of 31 bits. For channel coding, we adopt an LDPC code with an information word length of 24 and a codeword length of 49, denoted as LDPC(49, 24), resulting in a code rate close to 1/2. Subsequently, ECCT is used for algebraic block code decoding, which is capable of training on diverse error correction codes. The hyperparameter settings for ECCT training are detailed in Table 2. For comparative analysis, we select Deep Learning-Based Semantic Communication (DeepSC)[12], Universal Transformer (UT)[14], and UT

with quantization* as benchmark JSCC algorithms. Considering the subsequent signal-to-noise ratio (SNR) performance comparison, both algorithms are trained using mixed precision (i.e., float16), which, as discussed later, has a minimal negative impact on SNR computation. Key parameters used for training DeepSC and UT are also listed in Table 2. Besides, the traditional approach employs Huffman coding for source coding. Furthermore, bilingual evaluation understudy (BLEU)[51] and semantic similarity measured by BERT[52] are used to measure performance, as these metrics are widely recognized in natural language processing.

Most existing SemCom works evaluate the performance with respect to the SNR $= 10\log_{10}\left(E_{\mathrm{tb}}/N_0\right)$ dB, where E_{tb} denotes the energy associated with transmitting a single bit after source/channel coding and digital modulation, and N_0 represents the noise power spectral density. However, since different coding and modulation schemes across different communication methodologies result in varying numbers of bits transmitted over the physical channel, such a comparative metric of SNR ignores the differences in delivering different numbers of bits. Instead, referring to the total energy consumption E_{total} by sending $\mathrm{Num}_{\mathrm{unified}}$ bits through the physical channel in an LLM-based SSCC system, we propose a consistent definition of SNR in terms of an LLM-based SSCC reference baseline $\mathrm{SNR}_{\mathrm{unified}}$, as a function of the practically employed bits Num.

Table 2. Mainly used hyperparameters in the experiments

Model	Hyperparameter	Value
ECCT	Learning rate	10^{-4}
	Batch size	128
	Number of decoder layers	6
	Dimension of embedding	32
	Number of attention heads	8
DeepSC	Learning rate	10^{-4}
	Batch size	64
	Number of encoder/decoder layers	4
	Dimension of embedding	128
	Dimension of FFN	512
	Number of attention heads	8
UT	Learning rate	10^{-4}
	Batch size	64
	Number of encoder/decoder layers	3
	Dimension of embedding	128
	Dimension of FFN	1 024
	Number of attention heads	8

DeepSC: Deep Learning-Based Semantic Communication
ECCT: Error Correction Code Transformer
FFN: Feed Forward Network
LLM: Large Language Model
UT: Universal Transformer

Mathematically, this is expressed as:

$$\mathrm{SNR} = 10\log_{10}\left(\frac{E_{\mathrm{total}}}{N_0 \cdot \mathrm{Num}}\right) = 10\log_{10}\left(\frac{E_{\mathrm{total}}}{N_0 \cdot \mathrm{Num}_{\mathrm{unified}}} \times \frac{\mathrm{Num}_{\mathrm{unified}}}{\mathrm{Num}}\right) = \mathrm{SNR}_{\mathrm{unified}} + 10\log_{10}\left(\frac{\mathrm{Num}_{\mathrm{unified}}}{\mathrm{Num}}\right) \tag{9}$$

where $\mathrm{SNR}_{\mathrm{unified}}$ is used as an independent variable for aligning E_{total}, while for bit-oriented transmission (resp. float-based JSCC), Num denotes the number of bits (resp. float vectors) transmitted through the channel.

On the other hand, as mentioned in Section 1 and Ref. [33], deep learning-based JSCC systems extract the semantic feature of information to embed vectors in latent space, which is incompatible with digital communication systems. For JSCC methodologies like UT[14] and DeepSC[12], transmitting a float number certainly consumes far more energy than delivering a binary bit. In this case, if float16 is adopted, we can roughly assume it consumes an additional $10 \times \log_{10}(16) \approx 12.041$ dB. Hence, for the float-based JSCC methods, the unified evaluation metric is further modified to maintain a consistent energy consumption across different methodologies. In summary,

$$\mathrm{SNR} = \begin{cases} \mathrm{SNR}_{\mathrm{unified}} + 10\log_{10}\left(\frac{\mathrm{Num}_{\mathrm{unified}}}{\mathrm{Num}}\right) + 12.041, & \text{float based} \\ \mathrm{SNR}_{\mathrm{unified}} + 10\log_{10}\left(\frac{\mathrm{Num}_{\mathrm{unified}}}{\mathrm{Num}}\right), & \text{otherwise} \end{cases} \tag{10}$$

During evaluation, experiments are conducted for different schemes in terms of $\mathrm{SNR}_{\mathrm{unified}}$.

4.2 Numerical Results

In this section, we implement the GPT2-base model as a compressor and ECCT-complemented LDPC$(49, 24)$ as the error correction code, and compare it with DeepSC, UT, UT with quantization and the classical SSCC encompassing Huffman coding and ECCT (Fig. 5). The results demonstrate the superior performance of the proposed SSCC over the other three schemes. Similarly, we evaluate the performance under a Rayleigh fading channel in Fig. 6, where the results show that our system has a clear advantage in terms of the word-level BLEU score. However, in terms of semantic similarity, both the LLM-based and the traditional Huffman-based SSCC systems exhibit some disadvantages at lower SNRs, but still maintain a noticeable advantage at high SNRs.

In addition to presenting our key experimental results

* Compared to DeepSC and UT that directly transmit the encodes floats, UT with quantization maps the encoding results to a fixed number (30) of bits for transmission.

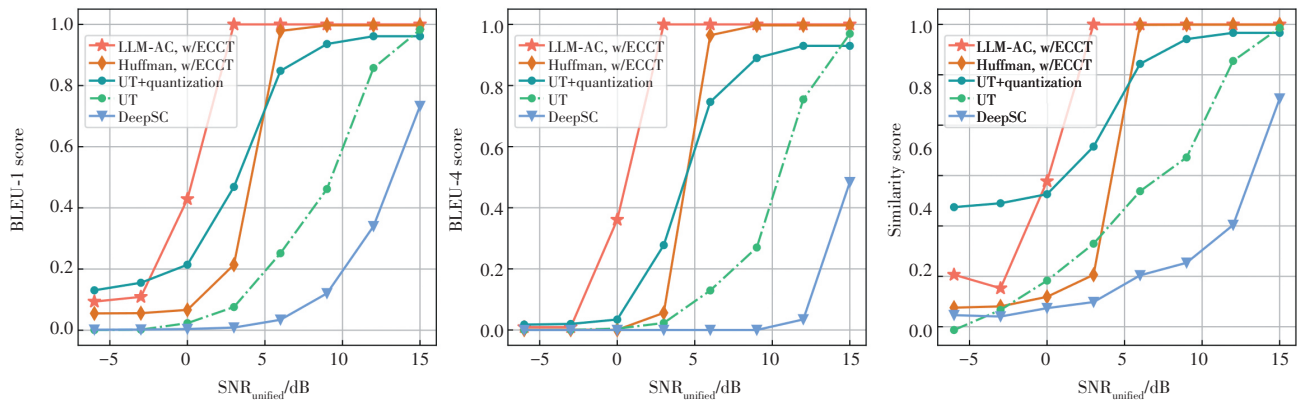


Figure 5. BLEU and similarity scores versus $SNR_{unified}$ are evaluated for the same number of transmitted symbols. The proposed LLM-based SSCC is compared with Huffman coding with LDPC(49, 24) in BPSK, DeepSC, UT, and UT with quantization under the AWGN channel

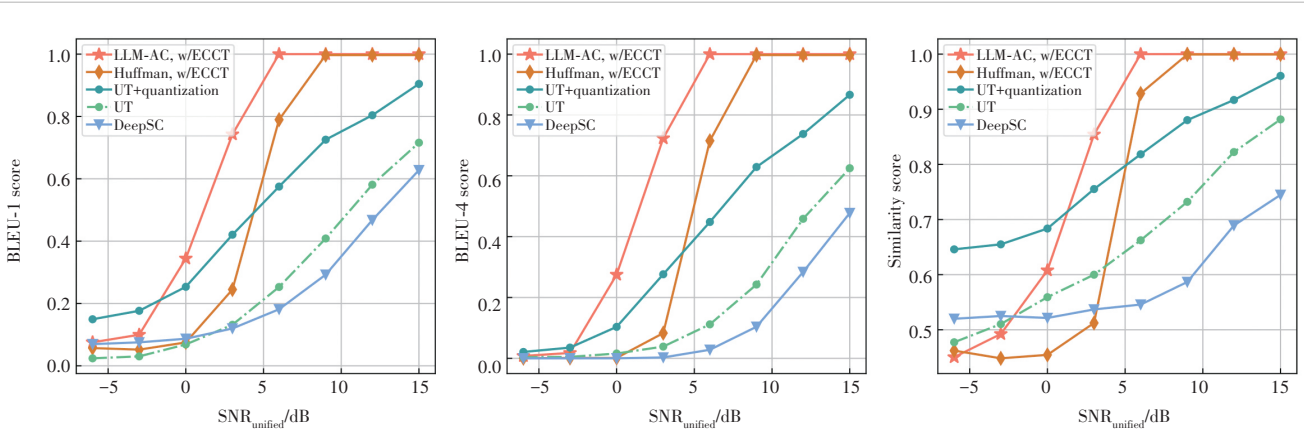


Figure 6. BLEU and similarity scores versus $SNR_{unified}$ are evaluated for the same number of transmitted symbols. The proposed LLM-based SSCC is compared with Huffman coding with LDPC(49, 24) in BPSK; DeepSC, UT, and UT with quantization trained under the Rayleigh fading channel

with $SNR_{unified}$ as the alignment metric, Fig. 7 provides performance comparisons using traditional SNR alignment, as well as the ratio of E_{total} used by different systems over our LLM-based solution. This illustrates the additional energy consumption of JSCC systems in achieving superior performance. Apparently, JSCC systems in SemCom achieve significant gains mainly due to the extra energy consumption.

Afterward, we validate the contributing effectiveness of ECCT[32] by comparing the performance of error correction codes with different coding rates under the same code length. Without loss of generality, the evaluation results based on LDPC under AWGN and Rayleigh fading channels are given in Fig. 8. Notably, while the work in Ref. [32] does not include Rayleigh channel results, inspired by the subsequent work on Denoising Diffusion Error Correction Codes (DDECC[53]), we extend ECCT to Rayleigh channels in a similar manner. It can be observed from Fig. 8 that compared to traditional LDPC decoding methods such as bit-flipping, ECCT provides consistent performance improvements. Furthermore, for error correction codes of the same length, lower coding rates demonstrate better recovery of noisy signals under the same SNR. More importantly, without ECCT, traditional algorithms struggle to decode noisy signals under Rayleigh channels effectively, and reducing the coding rate slightly improves the performance trivially. However, ECCT trained under the Rayleigh channel achieves as competitive

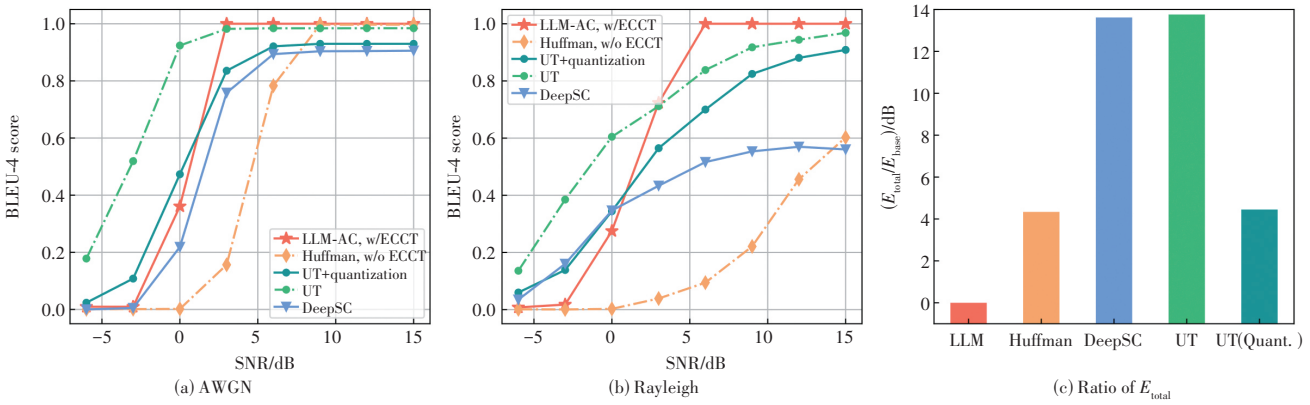


Figure 7. BLEU-4 score versus SNR is evaluated for the same number of transmitted symbols. The proposed LLM-based SSCC is compared with Huffman coding with LDPC(49, 24) in BPSK (without ECCT), DeepSC, UT, and UT with quantization trained under (a) AWGN and (b) Rayleigh fading channels; (c) shows the ratio of E_{total} among different systems

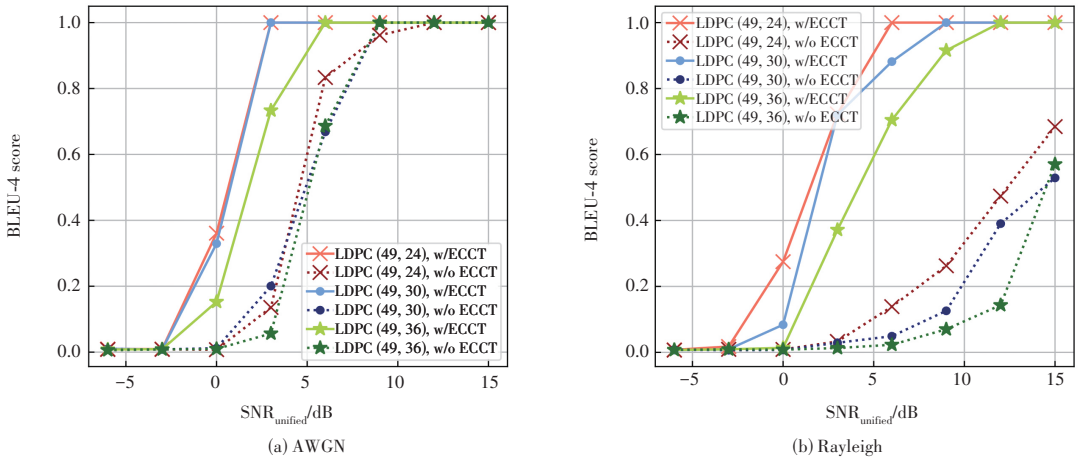


Figure 8. BLEU-4 score versus $SNR_{unified}$ for the same number of transmitted symbols, with different code rates using LDPC(49, 24)/ LDPC(49, 30)/ LDPC(49, 36) in BPSK, compared with the situations removing ECCT, under (a) AWGN and (b) Rayleigh fading channels

performance as that under the AWGN channel.

Considering the scaling law and emergent abilities of LLMs, we evaluate the performance by combining different models from the GPT2 family with ECCT-complemented LDPC channel coding (i. e., a high-rate LDPC(121, 110) code). Both the end-to-end SSCC performance in Fig. 9 and the compression rate in Fig. 10 indicate a notable performance improvement after adopting a model larger than GPT2. However, the performance difference among GPT2-medium, GPT2-large, and GPT2-XL is marginal. We hypothesize that while increasing the model size beyond a certain threshold contributes significantly to system performance, variations within a specific range of model scales yield diminishing returns. Furthermore, inspired by Ref. [54], the performance comparison with Zlib and static Huffman coding in Fig. 10 demonstrates that LLM-based arithmetic coding significantly outperforms traditional methods. Moreover, a scaling law is observed in the compression performance, which somewhat corroborates the findings of Ref. [54].

The experimental results presented in Table 3 further investigate the influence of the token block size on performance. It can be observed that at higher SNR levels, the performance generally improves as the block size increases, indicating that larger block sizes facilitate enhanced semantic

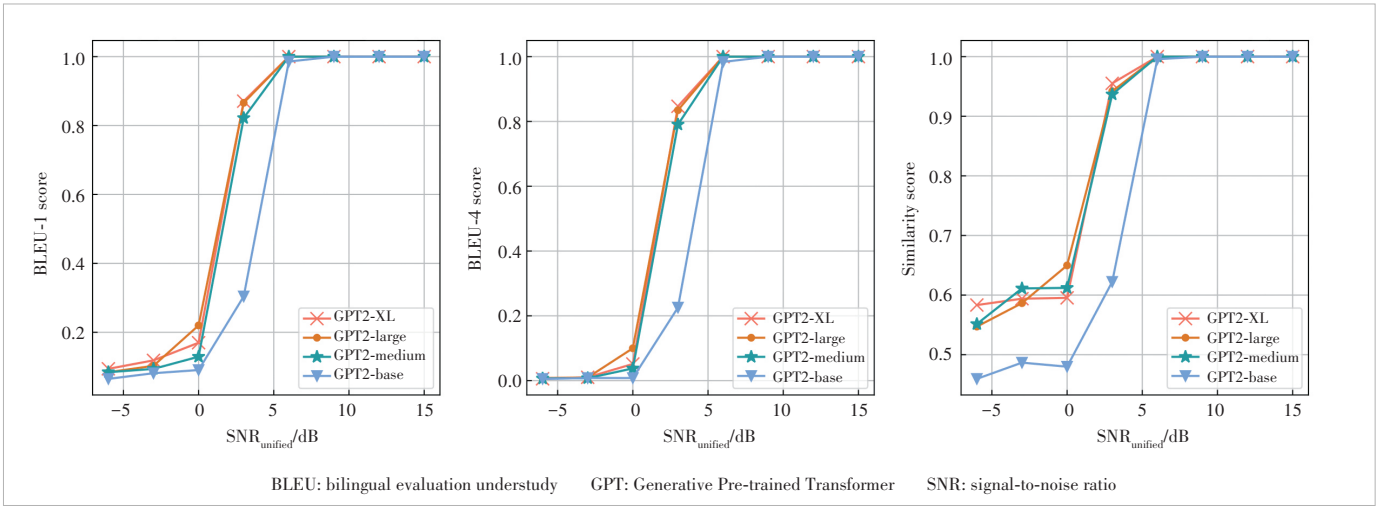


Figure 9. BLEU and similarity scores of models versus SNR$_{unified}$, with different parameter scales (GPT2, GPT2-medium, GPT2-large, GPT2-XL), using LDPC(121, 110) as the error correction code

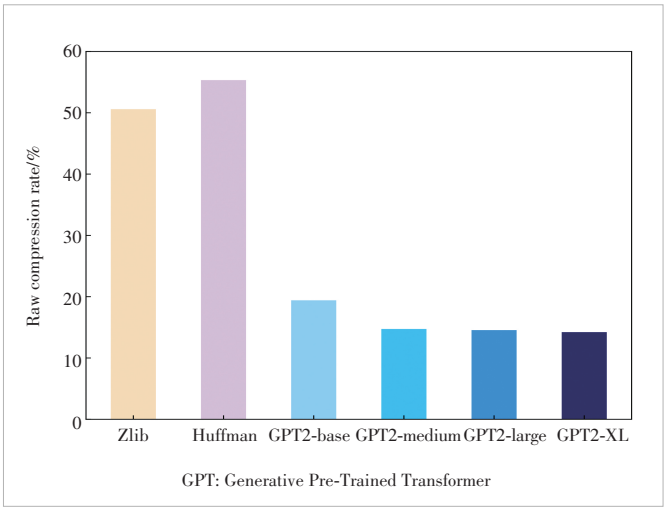


Figure 10. Compression rate comparison between traditional methods (Zlib and Huffman coding) and LLM-AC

Table 3. Influence of token block sizes on system performance during LLM-based arithmetic source encoding for SNR={−6, 0, 6}

Block size	Similarity			BLEU-1			BLEU-4		
	−6	0	6	−6	0	6	−6	0	6
16	**0.770 8**	0.915 7	0.999 3	**0.197 5**	**0.645 2**	0.987 7	**0.007 2**	**0.508 1**	0.983 0
32	0.712 3	**0.935 9**	0.998 4	0.172 5	0.584 2	0.978 7	0.005 5	0.466 6	0.969 4
64	0.700 1	0.893 8	0.999 9	0.116 0	0.580 1	0.996 9	0.001 8	0.427 0	0.992 2
128	0.758 7	0.857 3	**0.999 9**	0.183 1	0.434 4	**0.999 9**	0.003 8	0.252 9	**0.999 9**

BLEU: bilingual evaluation understudy SNR: signal-to-noise ratio
LLM: Large Language Model

preservation due to their ability to capture more contextual information. However, at lower SNR levels, the performance declines with an increase in the block size, suggesting that smaller blocks may be more resilient to avoid cumulative source decoding errors in these challenging scenarios.

5 Conclusions and Discussions

In this paper, we present a comprehensive analysis and evaluation of SSCC, with a comprehensive comparison to JSCC in the context of SemCom. Our proposed SSCC framework, which integrates LLMs for source coding and ECCT for enhanced channel coding, demonstrates significant performance improvements over JSCC in terms of recovery performance at both the word and semantic levels under both AWGN and Rayleigh fading channels. This highlights the potential effectiveness of SSCC in information transmission. In particular, through extensive experiments, we validate the strong compressive capability of LLMs to eliminate redundancy in text and the robustness of ECCT in enhancing decoding reliability under various channel conditions. In a word, separate source channel coding is still what we need.

Nevertheless, despite the validated performance superiority of SSCC, there remain several important issues worthy of further clarification and investigation.

1) The performance evaluation of text transmission sounds inspiring. The proposed SSCC framework is channel-agnostic, while given the well-known generality issues, the DNN-based JSCC faces a performance decline when the channel changes significantly. However, an extension to image transmission can be more challenging, and several issues like sequential tokenization require effective solutions. In this regard, potential solutions can incorporate patch division from Vision Transformer (ViT) to replace text tokenization, thereby segmenting images into semantic units for encoding. Consequently, the LLM-AC text predictor can be transformed into a probability modeler for image patches. Furthermore, the iterative decoding of ECCT can mitigate the error propagation issues in traditional JSCC, which is particularly crucial for multimedia transmission with high-

fidelity requirements. On the other hand, our experimental experience indicates the accuracy of channel coding is of vital importance for end-to-end performance. Hence, we only consider a relatively low, fixed code rate here. However, systematic tuning of the code rate is also a worthwhile direction for future research.

2) This paper only considers the classical JSCC design, while ignoring the latest quantization and digital modulation techniques that have emerged in the development of JSCC. For example, Refs. [9] and [10] show that utilizing a sparsity module to quantize the image embedding can yield significant performance gain. However, Refs. [9] and [10] have not compared their approaches with the remarkable capabilities of LLMs, and thus it remains unclear whether these amendments would enable JSCC to surpass LLM-based SSCC in a fair comparison. Nevertheless, given the inspiring results in this paper, there is no doubt that SSCC should be carefully improved rather than dismissed.

3) What we have to acknowledge is that integrating LLMs into the SSCC framework requires substantial computational resources for both encoding and decoding processes. However, we currently leverage pre-trained LLMs, which possess inherent generalization capabilities and can handle a broad range of natural language datasets. This contrasts with JSCC methods, which often rely on training with specific datasets to achieve superior performance. If a specific dataset is employed, we can explore the possibility of model distillation. By utilizing a Transformer model with significantly fewer parameters while retaining the LLM's tokenizer and performing self-supervised training on the target dataset, we can substantially reduce computational overhead while maintaining reasonable performance. We will further investigate model distillation in future work.

4) The discussions on JSCC are limited to the scenario to recover the semantics as accurately as possible. For SemCom[1], effectiveness-level or pragmatic communications may target at accomplishing different tasks under remotely controlled, noisy environments, rather than simple recovery of accurate semantics. In such cases, the underlying philosophy of JSCC may offer unique advantages.

5) Extensive works have been conducted to improve the performance of model-free decoders. For example, Ref. [55] proposes a systematic and double mask eliminating the difficulty of identifying the optimal parity-check matrix (PCM) from numerous candidatesfrom the same code. For performance enhancement on moderate code-length decoding, U-ECCT is proposed in Ref. [56] inspired by U-Net, while in Ref. [53], the Denoising Diffusion Probabilistic Model (DDPM)[57] is employed to model the transmission over channels as a diffusion process. Furthermore, a foundation model for channel codes is proposed in Ref. [58] for application to unseen codes. Therefore, these recent works are worthy to be evaluated in the SSCC framework.

References

[1] LU Z L, LI R P, LU K. Semantics-empowered communication: a tutorial-cum-survey [J]. IEEE communications surveys and tutorials, 2024, 26 (1): 41 – 79. DOI: 10.1109/COMST.2023.333334
[2] KURKA D B, GÜNDÜZ D. DeepJSCC-f: deep joint source-channel coding of images with feedback [J]. IEEE journal on selected areas in information theory, 2020, 1(1): 178 – 193. DOI: 10.1109/JSAIT.2020.2987203
[3] BAO Z C, LIANG H T, DONG C, et al. MDVSC: wireless model division video semantic communication for 6G [C]//Proc. IEEE Globecom Workshops (GC Wkshps). IEEE, 2023: 1572 – 1578. DOI: 10.1109/GCWkshps58843.2023.10464666
[4] JIA Y J, HUANG Z, LUO K, et al. Lightweight joint source-channel coding for semantic communications [J]. IEEE communications letters, 2023, 27(12): 3161 – 3165. DOI: 10.1109/LCOMM.2023.3329533
[5] LIU S C, GAO Z, CHEN G J, et al. Transformer-based joint source channel coding for textual semantic communication [C]//Proc. IEEE/CIC International Conference on Communications in China (ICCC). IEEE, 2023: 1 – 6. DOI: 10.1109/ICCC57788.2023.10233424
[6] LIU X Y, HUANG Z, ZHANG Y L, et al. CNN and attention-based joint source channel coding for semantic communications in WSNs [J]. Sensors, 2024, 24(3): 957. DOI: 10.3390/s24030957
[7] LU Z L, LI R P, LEI M, et al. Self-critical alternate learning based semantic broadcast communication [J]. IEEE transactions on communications, 2024: 1. DOI: 10.1109/tcomm.2024.3487513
[8] TONG W J, LIU F F, SUN Z F, et al. Image semantic communications: an extended rate-distortion theory based scheme [C]//Proc. IEEE Globecom Workshops (GC Wkshps). IEEE, 2022: 1723 – 1728. DOI: 10.1109/GCWkshps56602.2022.10008733
[9] TONG S Y, YU X X, LI R P, et al. Alternate learning based sparse semantic communications for visual transmission [C]//Proc. 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC). IEEE, 2023. DOI: 10.1109/pimrc56721.2023.10293971
[10] TONG S Y, YU X X, LI R P, et al. Alternate learning-based SNR-adaptive sparse semantic visual transmission [J]. IEEE transactions on wireless communications, 2025, 24: 1737 – 1752. DOI: 10.1109/TWC.2024.3512652
[11] WANG J, WANG S X, DAI J C, et al. Perceptual learned source-channel coding for high-fidelity image semantic transmission [C]//Proc. IEEE Global Communications Conference. IEEE, 2022: 3959 – 3964. DOI: 10.1109/GLOBECOM48099.2022.10001359
[12] XIE H Q, QIN Z J, LI G Y, et al. Deep learning enabled semantic communication systems [J]. IEEE transactions on signal processing, 2021, 69: 2663 – 2675. DOI: 10.1109/tsp.2021.3071210
[13] ZHANG W Y, BAI K Y, ZEADALLY S, et al. DeepMA: end-to-end deep multiple access for wireless image transmission in semantic communication [J]. IEEE transactions on cognitive communications and networking, 10(2): 387 – 402. DOI: 10.1109/tccn.2023.3326302
[14] ZHOU Q Y, LI R P, ZHAO Z F, et al. Semantic communication with adaptive universal transformer [J]. IEEE wireless communications letters, 2022, 11(3): 453 – 457. DOI: 10.1109/LWC.2021.3132067
[15] GOYAL M, TATWAWADI K, CHANDAK S, et al. DeepZip: lossless data compression using recurrent neural networks [C]//Proc. Data Compression Conference (DCC). IEEE, 2019. DOI: 10.1109/dcc.2019.00087
[16] BELLARD F. Lossless data compression with neural networks [EB/OL]. (2019-05-04)[2024-11-20]. https://bellard.org/nncp/nncp.pdf
[17] LIU Q, XU Y L, LI Z. DecMac: a deep context model for high efficiency arithmetic coding [C]//Proc. International Conference on Artificial Intelligence in Information and Communication (ICAIIC). IEEE, 2019. DOI: 10.1109/icaiic.2019.8668843
[18] GOYAL M, TATWAWADI K, CHANDAK S, et al. DZip: improved

general-purpose loss less compression based on novel neural network modeling [C]//Proc. Data Compression Conference (DCC). IEEE, 2021. DOI: 10.1109/dcc50243.2021.00023
[19] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proc. 31st International Conference on Neural Information Processing Systems. NIPS, 2017: 6000–6010
[20] HUANG C, XIE Y Q, JIANG Z Y, et al. Approximating human-like few-shot learning with GPT-based compression [EB/OL]. (2023-08-14) [2024-11-12]. https://arxiv.org/abs/2308.06942v1
[21] MITTU F, BU Y H, GUPTA A, et al. FineZip: pushing the limits of large language models for practical lossless text compression [EB/OL]. (2024-09-25)[2024-11-12]. https://arxiv.org/abs/2409.17141v1
[22] MAO Y, CUI Y F, KUO T W, et al. A fast transformer-based general-purpose lossless compressor [EB/OL]. (2022-03-30) [2024-11-12]. https://arxiv.org/abs/2203.16114v2
[23] NARASHIMAN S S, CHANDRACHOODAN N. AlphaZip: neural network-enhanced lossless text compression [EB/OL]. (2024-09-23) [2024-11-12]. https://arxiv.org/abs/2409.15046v1
[24] VALMEEKAM C S K, NARAYANAN K, KALATHIL D, et al. LLMZip: Lossless text compression using large language models [EB/OL]. (2023-06-06)[2024-11-12] https://arxiv.org/abs/2306.04050v2
[25] DELÉTANG G, RUOSS A, DUQUENNE P-A, et al. Language modeling is compression [EB/OL]. (2023-09-19)[2024-10-20]. https://arxiv.org/abs/2309.10668
[26] BOSE R C, RAY-CHAUDHURI D K. On a class of error correcting binary group codes [J]. Information and control, 1960, 3(1): 68–79. DOI: 10.1016/s0019-9958(60)90287-4
[27] GALLAGER R. Low-density parity-check codes [J]. IRE transactions on information theory, 1962, 8(1): 21–28. DOI: 10.1109/TIT.1962.1057683
[28] ARIKAN E. Channel polarization: a method for constructing capacity-achieving codes for symmetric binary-input memoryless channels [J]. IEEE transactions on information theory, 2009, 55(7): 3051–3073. DOI: 10.1109/TIT.2009.2021379
[29] NACHMANI E, BE'ERY Y, BURSHTEIN D. Learning to decode linear codes using deep learning [C]//Proc. 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton). IEEE, 2016: 341–346. DOI: 10.1109/ALLERTON.2016.7852251
[30] NACHMANI E, MARCIANO E, LUGOSCH L, et al. Deep learning methods for improved decoding of linear codes [J]. IEEE journal of selected topics in signal processing, 12(1): 119–131. DOI: 10.1109/jstsp.2017.2788405
[31] NACHMANI E, WOLF L. Hyper-graph-network decoders for block codes [C]//Proc. 33st International Conference on Neural Information Processing Systems. NIPS, 2019: 2329–2339
[32] CHOUKROUN Y, WOLF L. Error correction code transformer [C]//Proc. 36st International Conference on Neural Information Processing Systems. NIPS, 2022: 38695–38705
[33] HUANG J H, YUAN K, HUANG C, et al. D2-JSCC: digital deep joint source-channel coding for semantic communications [EB/OL]. (2024-03-12)[2024-11-20]. https://arxiv.org/abs/2403.07338v3
[34] JIANG P W, WEN C K, YI X P, et al. Semantic communications using foundation models: design approaches and open issues [J]. IEEE wireless communications, 2024, 31(3): 76–84. DOI: 10.1109/MWC.002.2300460
[35] LIANG C S, DU H Y, SUN Y, et al. Generative AI-driven semantic communication networks: architecture, technologies and applications [J]. IEEE transaction on cognitive communications and networking, 2024, early access. DOI: 10.1109/TCCN.2024.3435524
[36] JIANG F B, PENG Y B, DONG L, et al. Large AI model-based semantic communications [J]. IEEE wireless communications, 31(3): 68–75. DOI: 10.1109/mwc.001.2300346
[37] GRASSUCCI E, BARBAROSSA S, COMMINIELLO D. Generative semantic communication: diffusion models beyond bit recovery [EB/OL]. (2023-06-07)[2024-11-12]. https://arxiv.org/abs/2306.04321v1
[38] CHANG M K, HSU C T, YANG G C. GenSC: generative semantic communication systems using BART-like model [J]. IEEE communications letters, 2024, 28(10): 2298–2302. DOI: 10.1109/LCOMM.2024.3450309
[39] GUO S S, WANG Y H, LI S J, et al. Semantic importance-aware communications using pre-trained language models [J]. IEEE communications letters, 2023, 27(9): 2328–2332. DOI: 10.1109/LCOMM.2023.3293805
[40] XIE H Q, QIN Z J, TAO X M, et al. Toward intelligent communications: large model empowered semantic communications [J]. IEEE communications magazine, 2025, 63(1): 69–75. DOI: 10.1109/MCOM.001.2300807
[41] QIAO L, MASHHADI M B, GAO Z, et al. Latency-aware generative semantic communications with pre-trained diffusion models [EB/OL]. (2024-03-05)[2024-11-12]. https://arxiv.org/abs/2403.17256v2
[42] JIANG F B, DONG L, PENG Y B, et al. Large AI model empowered multimodal semantic communications [J]. IEEE communications magazine, 2025, 63(1): 76–82. DOI: 10.1109/mcom.001.2300575
[43] YANG W T, XIONG Z H, MAO S W, et al. Rethinking generative semantic communication for multi-user systems with large language models [EB/OL]. (2024-08-16) [2024-11-12]. https://arxiv. org/abs/2408.08765v3
[44] SHANNON C E. A mathematical theory of communication [J]. Bell system technical journal, 1948, 27(3): 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x
[45] RISSANEN J J. Generalized kraft inequality and arithmetic coding [J]. IBM journal of research and development, 1976, 20(3): 198–203. DOI: 10.1147/rd.203.0198
[46] PASCO R. Source coding algorithms for fast data compression (Ph.D. Thesis abstr.) [J]. IEEE transactions on information theory, 1977, 23 (4): 548. DOI: 10.1109/TIT.1977.1055739
[47] HOWARD P G, VITTER J S. Arithmetic coding for data compression [J]. Proceedings of the IEEE, 1994, 82(6): 857–865. DOI: 10.1109/5.286189
[48] BENNATAN A, CHOUKROUN Y, KISILEV P. Deep learning for decoding of linear codes: a syndrome-based approach [C]//Proc. IEEE International Symposium on Information Theory (ISIT). IEEE, 2018: 1595–1599. DOI: 10.1109/ISIT.2018.8437530
[49] KOEHN P. Europarl: a parallel corpus for statistical machine translation [C]//Proc. Machine Translation Summit. International Association for Machine Translation, 2005: 79–86
[50] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners [EB/OL]. [2024-10-20]. https://cdn. openai. com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
[51] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation [C]//Proc. 40th Annual Meeting on Association for Computational Linguistics. USAACL, 2001. DOI: 10.3115/1073083.1073135
[52] DEVLIN J, CHANG M-W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]//Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Association for Computational Linguistics, 2019: 4171–4186. DOI: 10.18653/v1/N19-1423
[53] CHOUKROUN Y, WOLF L. Denoising diffusion error correction codes [EB/OL]. (2022-09-16) [2024-11-12]. https://arxiv. org/abs/2209.13533v1
[54] HUANG Y Z, ZHANG J H, SHAN Z F, et al. Compression represents intelligence linearly [EB/OL]. (2024-04-15)[2024-11-12]. https://arxiv.org/abs/2404.09937v2
[55] PARK S J, KWAK H Y, KIM S H, et al. How to mask in error correction code transformer: systematic and double masking [EB/OL]. (2023-08-16)[2024-11-12]. https://arxiv.org/abs/2308.08128v2

[56] NGUYEN D T, KIM S. U-shaped error correction code transformers [J]. IEEE transactions on cognitive communications and networking, 2024: 1. DOI: 10.1109/tccn.2024.3482349
[57] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models [C]//Proc. 34th International Conference on Neural Information Processing Systems. NIPS, 2020: 6840 – 6851. DOI: 10.48550/arXiv.2006.11239
[58] CHOUKROUN Y, WOLF L. A foundation model for error correction codes [C]//12th International Conference on Learning Representations. ICLR, 2024. DOI: 10.48550/arXiv.2405.04050

Appendix 1: Pseudo code of finite precision arithmetic codec

Appendix A: Pseudo code for encoder

Algorithm 1 Finite-precision arithmetic encoding

Require: N_k: Current number of emitted bits m_{N_k}

Require: $p_{\text{cum}}\left(D_i|\boldsymbol{t}_{1:k}\right)$: Cumulative probability of token $t_{k+1}=D_i\in\mathcal{D}$ given first k tokens

Require: l_k, u_k: Current interval determined by the first k tokens

Require: ε_k: Number of scaling bits

Initialization:
$N_{k+1}\leftarrow N_k$
$l_{k+1}\leftarrow l_k+\left(u_k-l_k\right)p_{\text{cum}}\left(D_{i-1}|\boldsymbol{t}_{1:k}\right)$ // If $k=0$, use $p_{\text{cum}}\left(D_{i-1}\right)$
$h_{k+1}\leftarrow l_k+\left(u_k-l_k\right)p_{\text{cum}}\left(D_i|\boldsymbol{t}_{1:k}\right)$ // If $k=0$, use $p_{\text{cum}}\left(D_i\right)$
$\varepsilon_{k+1}\leftarrow\varepsilon_k$
Scaling:
while any of the scaling conditions is met **do**
  if $u_{k+1}<0.5$ **then**
    // Scaling 1
    $l_{k+1}, u_{k+1}\leftarrow 2l_{k+1}, 2u_{k+1}$
    $m_{N_{k+1}+1}\leftarrow 0$ // Emit one bit's 0
    $m_{N_{k+1}+2:N_{k+1}+1+\varepsilon_{k+1}}\leftarrow 1$ // Emit ε_{k+1} bits' 1
    $N_{k+1}\leftarrow N_{k+1}+1+\varepsilon_{k+1}$
    $\varepsilon_{k+1}\leftarrow 0$
  else if $l_{k+1}\geqslant 0.5$ **then**
    // Scaling 2
    $l_{k+1}, u_{k+1}\leftarrow 2\left(l_{k+1}-0.5\right), 2\left(u_{k+1}-0.5\right)$
    $m_{N_{k+1}+1}\leftarrow 1$ // Emit one bit's 1
    $m_{N_{k+1}+2:N_{k+1}+1+\varepsilon_{k+1}}\leftarrow 0$ // Emit ε_{k+1} bits' 0
    $N_{k+1}\leftarrow N_{k+1}+1+\varepsilon_{k+1}$
    $\varepsilon_{k+1}\leftarrow 0$
  else if $0.25\leqslant l_{k+1}<0.5\leqslant u_{k+1}<0.75$ **then**
    // Scaling 3
    $l_{k+1}, u_{k+1}\leftarrow 2\left(l_{k+1}-0.25\right), 2\left(u_{k+1}-0.25\right)$
    $\varepsilon_{k+1}\leftarrow\varepsilon_{k+1}+1$
  end if
end while
return $N_{k+1}, \boldsymbol{m}_{N_k+1:N_{k+1}}$ // Updated emitted bits

Appendix B: Pseudo code for decoder

Algorithm 2 Finite-precision arithmetic decoding

Require: K_n: Current number of decoded tokens

Require: $p_{\text{cum}}\left(D_i|\boldsymbol{t}_{1:K_n}\right)$: Cumulative probability of token $t_{K_{n+1}+1}=D_i\in\mathcal{D}$ given first K_n tokens

Require: l_n, u_n: Current interval determined by the first n bits

Require: l_{K_n}, u_{K_n}: Interval of sequence $t_{1:K_n}$ that has been decoded

Initialization:
$K_{n+1}\leftarrow K_n$
$l_{K_{n+1}}, h_{K_{n+1}}\leftarrow l_{K_n}, h_{K_n}$
if the $(n+1)$-th bit $m_{n+1}=0$ **then**
  $l_{n+1}, h_{n+1}\leftarrow l_n, \frac{1}{2}\left(l_n+h_n\right)$
else
  $l_{n+1}, h_{n+1}\leftarrow \frac{1}{2}\left(l_n+h_n\right), h_n$
end if
while Not End-of-Sentence symbol **do**
  Search:
  Find $D_i\in\mathcal{D}$ such that:
    $L=l_{K_{n+1}}+\left(u_{K_{n+1}}-l_{K_{n+1}}\right)p_{\text{cum}}\left(D_{i-1}|t_{1:K_n}\right)$ // If $K_n=0$, use $p_{\text{cum}}\left(D_{i-1}\right)$
    $U=l_{K_{n+1}}+\left(u_{K_{n+1}}-l_{K_{n+1}}\right)p_{\text{cum}}\left(D_i|t_{1:K_n}\right)$ // If $K_n=0$, use $p_{\text{cum}}\left(D_i\right)$
    $L\leqslant l_{n+1}<u_{n+1}<U$ // i. e. current interval of the n-th bit is included in the interval of D_i
  if D_i exists **then**
    Update:
    $K_{n+1}\leftarrow K_{n+1}+1$
    $t_{K_{n+1}}\leftarrow D_i$//Output D_i to the token sequence t
    $l_{K_{n+1}}, u_{K_{n+1}}\leftarrow L, U$
    Scaling: Similar to the Scaling in Algorithm 1
    Go to Search
  else
    return $K_{n+1}, t_{K_n+1:K_{n+1}}$
  end if
end while
return $K_{n+1}, t_{K_n+1:K_{n+1}}$ // Updated decoded tokens

Appendix 2: Two key modules of ECCT

Appendix C: Positional reliability encoding

For the channel output $\boldsymbol{y}$, the positional reliability encoding transforms each dimension of $\tilde{\boldsymbol{y}}$ into a high d dimensional embedding ϕ, which enriches the information of input embedding vectors and replaces $\tilde{\boldsymbol{y}}$ as the input of ECCT. The transformation is defined by

$$\phi_i = \begin{cases} |\boldsymbol{y}_i| \boldsymbol{W}_i, & \text{if } i \leqslant N \\ \text{bin\_to\_sign}\left(\text{syn}\left(\boldsymbol{y}_{i-N+1}\right)\right)\boldsymbol{W}_i, & \text{otherwise} \end{cases} \tag{11},$$

where $\{\boldsymbol{W}_i \in \mathbb{R}^d\}_{i=1}^{2N-K}$ denotes the learnable embedding matrix representing the bit's position-dependent one-hot encoding. The encoding method corresponds to the input reliability and is positional, since unreliable information of low magnitude would collapse to the origin, while the syndrome scales negatively. Hence, it is termed positional reliability encoding.

Appendix D: Code-aware self-attention

The code-aware attention mask mechanism aims to integrate code-specific sparse marks that incorporate the inherent structural characteristics of their respective PCM as the domain knowledge. Given a codeword defined by the generator matrix $\boldsymbol{G}$ and parity check matrix $\boldsymbol{H}$, the attention mask is defined by $\boldsymbol{g}(\boldsymbol{H}):\{0,1\}^{(n-k)\times n} \rightarrow \{-\infty, 0\}^{(2n-k)\times(2n-k)}$, the construction of which is shown in Algorithm 3. Then, the code-aware self-attention mechanism could be represented as

$$A_H(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T + \boldsymbol{g}(\boldsymbol{H})}{\sqrt{d}}\right)\boldsymbol{V} \tag{12},$$

where $\boldsymbol{Q}$, $\boldsymbol{K}^*$ and $\boldsymbol{V}$ denote the query, key and value in self-attention. During the implementation, the code-aware attention mask mechanism is used as an enhancement of the multi-head-attention module in the classical Transformer architecture.

Algorithm 3 Pseudo code of building the attention mask

Require: parity-check matrix $\boldsymbol{H}$ of error correction code $C_e(N,K)$

mask ← eye$(2N-K)$
for $i = 1, 2, \cdots, N-K$ **do**
  idx ← where$(\boldsymbol{H}[i] == 1)$
  for j in idx **do**
    mask$[N+i, j]$, mask$[j, N+i]$ ← 1
    for l in idx **do**
      mask$[j, l]$, mask$[l, j]$ ← 1
    end for
  end for
end for
mask ← $-\infty(\neg$mask$)$
return mask // Output attention mask $\boldsymbol{g}(\boldsymbol{H})$

Biographies

REN Tianqi received his BE degree in electronic science and technology from Zhejiang University, China in 2024. He is currently pursuing his ME degree in electronic and information engineering with Zhejiang University. His research interests include application of large language models in communication scenarios and semantic communications.

LI Rongpeng (lirongpeng@zju.edu.cn) is currently an associate professor with the College of Information Science and Electronic Engineering, Zhejiang University, China. He was a research engineer with the Wireless Communication Laboratory, Huawei Technologies Co., Ltd. from August 2015 to September 2016. He was a visiting scholar with the Department of Computer Science and Technology, University of Cambridge, UK from February 2020 to August 2020. His research interest currently focuses on networked intelligence for communications evolving (NICE). He received the Wu Wenjun Artificial Intelligence Excellent Youth Award in 2021. He serves as an Editor for *China Communications*.

ZHAO Mingmin received his BEng and PhD degrees in information and communication engineering from Zhejiang University, China in 2012 and 2017, respectively. From December 2015 to August 2016, he was a visiting scholar with the Department of Electrical and Computer Engineering, Iowa State University, USA. From July 2017 to July 2018, he was a research engineer with Huawei Technologies Co., Ltd. He is currently a lecturer with the College of Information Science and Electronic Engineering, Zhejiang University. Since May 2019, he has been a visiting scholar with the Department of Electrical and Computer Engineering, National University of Singapore. His research interests include channel coding, algorithm design and analysis for advanced MIMO, cooperative communication, and machine learning for wireless communications.

CHEN Xianfu received his PhD degree (with Hons.) from Zhejiang University, China in 2012. In 2012, he joined the VTT Technical Research Centre of Finland, as a research scientist and as a senior scientist from 2013 to 2023. He is currently a chief research engineer with the Shenzhen CyberAray Network Technology Co., Ltd., China. His research interests include various aspects of wireless communications and networking, with emphasis on human-level and artificial intelligence for resource awareness in next-generation communication networks. Dr. CHEN was the recipient of the 2021 IEEE Communications Society Outstanding Paper Award and the 2021 IEEE Internet of Things Journal Best Paper Award. He is an editor of *IEEE Open Journal of the Communications Society*, an academic editor of *Wireless Communications and Mobile Computing*, and an associate editor of *China Communications*.

LIU Guangyi received his PhD degree from Beijing University of Posts and Telecommunications, China in 2006. He is currently the chief scientist of 6G in China Mobile Communication Corporation (CMCC), the founding member and

* It is worthwhile to point that $\boldsymbol{K}$ distinguishes from K, which represents the length of the error correction code in the previous text.

the co-chair of the 6G Alliance of Network AI, and the vice-chair of the THz Industry Alliance in China and the Wireless Technology Working Group of IMT-2030 (6G) Promotion Group supported by Ministry of Information and Industry Technology of China. He has been leading the 6G research and development with CMCC since 2018. He has led the Research and Development of 4G's evolution and 5G in CMCC from 2006 to 2020. He has acted as a Spectrum Working Group Chair and the Project Coordinator of LTE Evolution and 5G eMBB in the Global TD-LTE Initiative from 2013 to 2020 and led the industrialization and globalization of TD-LTE evolution and 5G eMBB.

YANG Yang is a professor with the IoT Thrust, the Director of the Research Center for the Digital World with Intelligent Things (DOIT), and the associate vice-president for Teaching and Learning with The Hong Kong University of Science and Technology (Guangzhou), China. He is also an adjunct professor with the Department of Broadband Communication at Peng Cheng Laboratory, the chief scientist of IoT with Terminus Group, and a senior consultant for Shenzhen Smart City Technology Development Group, China. His research interests include multi-tier computing networks, 5G/6G systems, AIoT technologies, intelligent services and applications, and advanced wireless testbeds. He has been the chair of the Steering Committee of the Asia-Pacific Conference on Communications (APCC) from 2019 to 2021. Currently, he is serving the IEEE Communications Society as the chair for the 5G Industry Community and chair for the Asia Region at Fog/Edge Industry Community. He is a fellow of IEEE.

ZHAO Zhifeng received his BE degree in computer science, ME degree in communication and information systems, and PhD degree in communication and information systems from the PLA University of Science and Technology, China in 1996, 1999, and 2002, respectively. From 2002 to 2004, he acted as a post-doctoral researcher with Zhejiang University, China, where his studies focused on multimedia next-generation networks (NGNs) and softswitch technology for energy efficiency. Currently, he is with the Zhejiang Lab as the Chief Engineering Officer. His research areas include software-defined networks (SDNs), wireless networks in 6G, computing networks, and collective intelligence. He is the Symposium Co-Chair of ChinaCom 2009 and 2010. He is the TPC Co-Chair of the 10th IEEE International Symposium on Communication and Information Technology (ISCIT 2010).

ZHANG Honggang is a professor with the Faculty of Data Science, City University of Macau, China. He was the founding Chief Managing Editor of *Intelligent Computing*, a Science Partner Journal, and a professor with the College of Information Science and Electronic Engineering, Zhejiang University, China. He was an Honorary Visiting Professor with the University of York, UK, and an International Chair Professor of Excellence with the Université Européenne de Bretagne and Supélec, France. His research interests include cognitive radio networks, semantic communications, green communications, machine learning, artificial intelligence, intelligent computing, and the Internet of Intelligence. He is a co-recipient of the 2021 IEEE Communications Society Outstanding Paper Award and the 2021 IEEE Internet of Things Journal Best Paper Award. He was the leading guest editor for the special issues on green communications of the *IEEE Communications Magazine*. He is the associate editor-in-chief of *China Communications*. He is a fellow of IEEE.

Exploration of NWDAF Development Architecture for 6G AI-Native Networks

HE Shiwen[1,2], PENG Shilin[1], DONG Haolei[1], WANG Liangpeng[2], AN Zhenyu[2]

(1. School of Computer Science and Engineering, Central South University, Changsha 410083, China;
2. Purple Mountain Laboratories, Nanjing 210096, China)

DOI: 10.12142/ZTECOM.202501006

https://kns.cnki.net/kcms/detail/34.1294.TN.20250305.1750.002.html, published online March 6, 2025

Manuscript received: 2024-12-15

Abstract: Artificial intelligence (AI)-native communication is considered one of the key technologies for the development of 6G mobile communication networks. This paper investigates the architecture for developing the network data analytics function (NWDAF) in 6G AI-native networks. The architecture integrates two key components: data collection and management, and model training and management. It achieves real-time data collection and management, establishing a complete workflow encompassing AI model training, deployment, and intelligent decision-making. The architecture workflow is evaluated through a vertical scaling use case by constructing an AI-native network testbed on Kubernetes. Within this proposed NWDAF, several machine learning (ML) models are trained to make vertical scaling decisions for user plane function (UPF) instances based on data collected from various network functions (NFs). These decisions are executed through the Kubernetes API, which dynamically allocates appropriate resources to UPF instances. The experimental results show that all implemented models demonstrate satisfactory predictive capabilities. Moreover, compared with the threshold-based method in Kubernetes, all models show a significant advantage in response time. This study not only introduces a novel AI-native NWDAF architecture but also demonstrates the potential of AI models to significantly improve network management and resource scaling in 6G networks.

Keywords: 6G; AI-native; NWDAF; UPF scaling

Citation (Format 1): HE S W, PENG S L, DONG H L, et al. Exploration of NWDAF development architecture for 6G AI-native networks [J]. ZTE Communications, 2025, 23(1): 45 - 52. DOI: 10.12142/ZTECOM.202501006
Citation (Format 2): S. W. He, S. L. Peng, H. L. Dong, et al., "Exploration of NWDAF development architecture for 6G AI-native networks," *ZTE Communications*, vol. 23, no. 1, pp. 45 - 52, Mar. 2025. doi: 10.12142/ZTECOM.202501006.

1 Introduction

With the rapid advancement of mobile communication technologies, 6G communication networks have gained widespread attention as the next frontier in modern communication systems. A characteristic of 6G networks is the integration of artificial intelligence (AI) and network architecture, a concept termed AI-native communication[1]. The convergence of AI and communication technologies will transform network operations, enabling autonomous decision-making, dynamic resource management, and seamless adaptation to changing network conditions[2]. Meanwhile, AI-native communication has become a critical enabling technology for 6G, capturing the attention of academic and industrial communities worldwide. However, realizing AI-native communication functionality requires a robust foundation of supporting technologies to ensure flexibility, scalability, and efficiency for network operations. As an essential component of AI-native networks, the core network (CN) facilitates real-time data collection and intelligent resource scheduling[3]. By embedding AI directly into its architecture, the CN ensures seamless coordination among different network functions (NFs), enabling the dynamic adaptability and scalability essential for 6G networks[4]. Therefore, technologies such as network function virtualization (NFV), software-defined networking (SDN), and containerization have also been integrated into research on the intelligent evolution of the CN. NFV transforms dedicated NFs into virtualized software instances[5], reducing dependence on dedicated hardware and simplifying the AI-native network architecture. This also enables the seamless integration of AI models into network management, control, and optimization. SDN decouples the control and data planes, enabling centralized management and dynamic routing optimization, which supports real-time monitoring and adjustment of data flow paths, creating improved network conditions for AI models[6]. Additionally, containeriza-

This work was supported by the National Key Research and Development Program of China under Grant No. 2023YFE0200700, National Natural Science Foundation of China under Grant No. 62171474, and ZTE Industry-University-Institute Cooperation Funds under Grant No. IA20241014013.

tion provides modularity and lightweight deployment capabilities, allowing efficient deployment, management, and updating of AI models, thus providing flexible, efficient, and scalable infrastructure support for AI-native networks[7]. Although these techniques improve the flexibility, robustness, and efficiency of the CN, enabling better deployment and resource flexibility for AI-native networks, they alone are insufficient to fully realize the potential of AI-native capabilities[8]. Bridging this gap requires the incorporation of advanced data analysis and decision-making mechanisms[9], which are essential for achieving the autonomous intelligence envisioned in AI-native networks. The network data analytics function (NWDAF) addresses this critical gap by serving as an independent intelligent NF introduced in the CN, as part of the 5G standard protocol proposed by the 3rd Generation Partnership Project (3GPP) first[10]. By leveraging the flexibility and scalability enabled by NFV, SDN, and containerization, the NWDAF integrates advanced data analytics and machine learning techniques to effectively process and analyze large-scale datasets within the CN.

NFV enables the NWDAF to be dynamically deployed and scaled on demand, allowing real-time adjustment of computing resources based on network traffic. SDN facilitates multi-dimensional data collection from various NFs while enabling flexible traffic management. Additionally, containerization enhances the adaptability of the NWDAF by supporting rapid deployment, migration, and scaling across diverse environments, thereby improving the modularity and scalability of the data analysis process. These enable the NWDAF to provide intelligent data-driven decision-making support for network optimization and management. Through sophisticated data processing and predictive modeling, the NWDAF addresses critical challenges in modern networks, including resource allocation, load balancing, and fault recovery[11]. As a result, it facilitates the evolution of the CN from traditional reactive management to more intelligent and autonomous operational models, paving the way for more efficient, adaptive, and dynamic network management[12].

Despite the significant flexibility and intelligent potential demonstrated by the NWDAF through the integration of NFV, SDN, and containerization technologies, its efficient implementation still faces two core challenges. First, data collection serves as the foundation for both NWDAF and AI-native network implementation. Unlike traditional NFs, the NWDAF requires a comprehensive and real-time collection of network state data to support accurate analytics and decision-making[13]. This is particularly critical in scenarios such as user plane function (UPF) scaling, where real-time data on UPF load and resource availability must be continuously gathered to enable dynamic and efficient resource allocation. Second, AI-native capabilities demand the seamless embedding of AI into the network architecture rather than the standalone application of machine learning (ML) models[14]. Given the diversity of network scenarios, the NWDAF must include a dedicated function to manage the training, storage, and dynamic orchestration of various AI models, ensuring adaptability and scalability in meeting specific operational requirements. Specifically, in the context of UPF optimization, the NWDAF can facilitate seamless UPF scaling to accommodate varying traffic demands thus optimizing network performance. In light of these challenges, existing research has investigated various aspects of these challenges. For instance, MEKRACHE et al. proposed a microservice architecture for the NWDAF, employing a Long Short-Term Memory (LSTM) auto-encoder to detect abnormal traffic generated by user equipment (UE), utilizing the Milano dataset for network data analysis[15]. Furthermore, NISHA et al. proposed a network load prediction and anomaly detection method and tested it using various machine learning methods[16]. Their work employed a comprehensive dataset supporting 5G networks. However, these studies primarily rely on publicly available or simulated datasets, focusing on optimizing network performance for a specific scenario, without implementing the complete pipeline from data collection to data analysis. SEVGICAN et al. proposed a system for intelligent network analytics using ML techniques, comparing the performance of multiple machine learning models[17]. MANIAS et al. proposed a prototype system for the NWDAF within the CN, employing data-driven techniques and unsupervised learning to analyze NF interactions[18]. ZHANG et al. applied fair federated learning (FL) to the 3GPP-standard NWDAF architecture, integrating a multi-task ML model for anomaly traffic detection across different types of user devices[19]. Most existing studies focus on singular optimization tasks, often overlooking the comprehensive management of the model lifecycle. As a result, critical aspects such as real-time data collection, dynamic model training, and model orchestration across various scenarios remain underexplored. This limitation significantly hinders the practicality and generalization capability of models in real-world network environments. To address these challenges, this paper makes the following contributions:

1) An AI-native NWDAF architecture is designed to integrate the functionalities for data collection and management, providing a solution for complex data processing in 6G network environments.

2) A specialized module for model training and management is designed to enable dynamic adaptation of AI models to diverse scenarios, effectively meeting the intelligent management requirements of 6G networks.

3) An AI-native network testbed is constructed to evaluate the designed architecture functionalities, including data collection, model training, and management. Meanwhile, the UPF scaling scenarios are adopted to validate the feasibility and effectiveness of the proposed architecture.

The rest of the paper is structured as follows: Section 2 introduces the AI-native NWDAF architecture. Section 3 discusses system deployment and experimental validation.

Section 4 concludes the paper and outlines directions for future research.

2 System Architecture

To enhance the intelligent capabilities of the 6G CN, this paper proposes a systematic approach through a novel NWDAF architecture. As shown in Fig. 1, the proposed system architecture comprises two layers: the infrastructure layer and the NWDAF service layer. The infrastructure layer is implemented on Kubernetes, providing functionalities for network monitoring and optimization. It detects the network environment, orchestrates network instances, and provides network data to NWDAF consumers for optimization through intelligent algorithms. The NWDAF service layer contains required NFs to implement the CN, with each NF independently deployed using the containerization technology. The NWDAF functional architecture consists of two key components: data collection and management (DCM) and model training and management (MTM). DCM is responsible for data collection, storage, and distribution, ensuring a continuous supply of high-quality data for the MTM. MTM focuses on model training and analysis, while managing various ML models to adapt to specific scenarios. The proposed architecture encompasses the entire process from data collection to model training and application, enabling the system to achieve autonomous decision-making. Compared to traditional network architectures or those with externally attached AI components, it provides enhanced flexibility and intelligence.

2.1 Data Collection and Management

DCM serves as a fundamental component of the proposed architecture, functioning as the basis of model training and analysis. This module encompasses three functions: data collection, data storage, and data distribution. NWDAF collects data from multiple NFs, including the access and mobility management function (AMF), session management function (SMF), and UPF, which is subsequently stored for subsequent analysis and utilization.

To achieve efficient data collection and management, the system leverages Prometheus as the data collection framework. As illustrated in Fig. 2, each NF consists of two modules: the

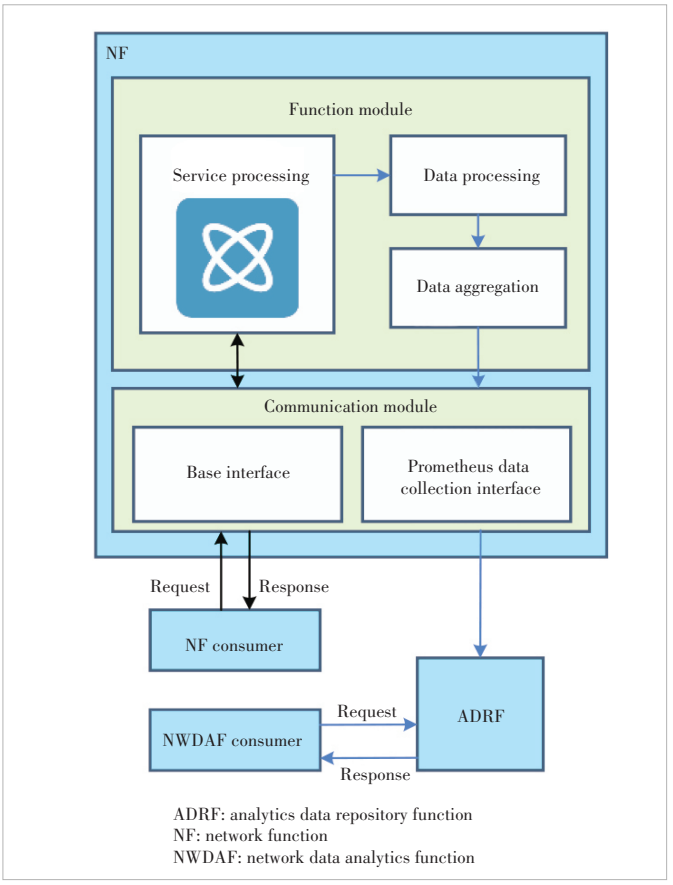


Figure 2. Data collection and management

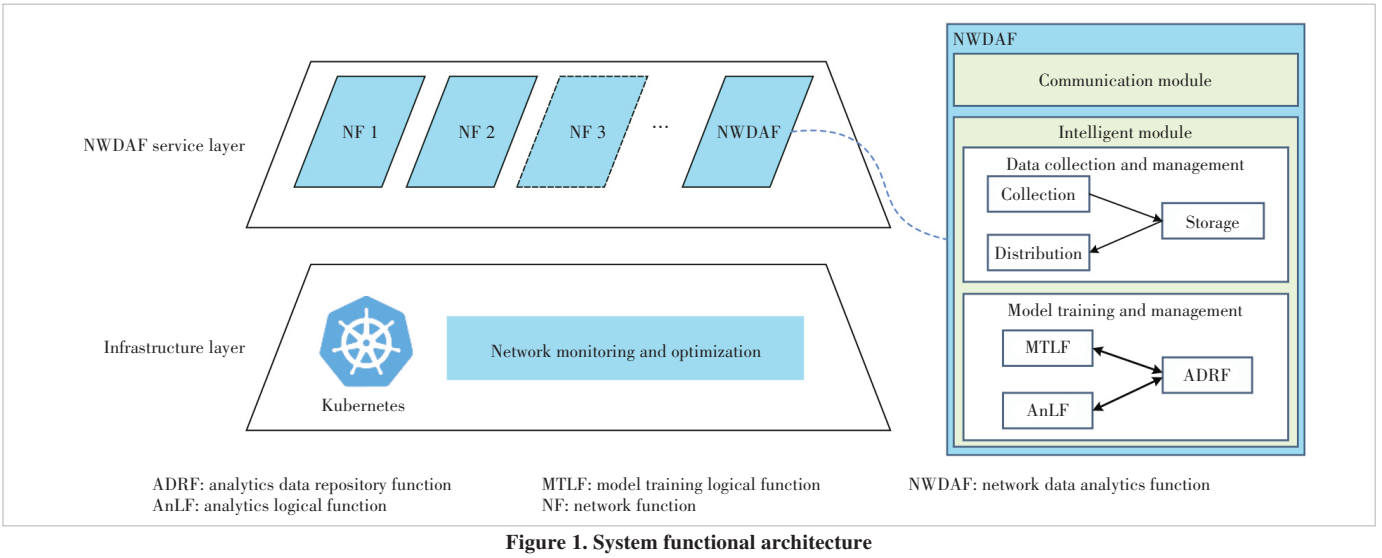


Figure 1. System functional architecture

function module and the communication module. The function module implements all NF's functionalities and handles various services. The communication module manages external communication, comprising the system's base interface and the customized Prometheus data collection interface. The NWDAF performs data collection through the Prometheus data collection interface, which operates in parallel with the signal processing tasks within the NFs. This approach bypasses the base interface, such as traditional N-interface communication between NFs. This design not only avoids interference with the service workflows of the NFs, but also ensures high efficiency and reliability in the data collection process. Specifically, when an NF consumer initiates a processing request, the NF executes the task within its function module and returns the results to the consumer without interruption. Meanwhile, the specified data are recorded and stored in a cache. Subsequently, the data are processed and aggregated within the NF, and then uploaded to the analytics data repository function (ADRF) database via the data collection interface.

The ADRF database is implemented through Prometheus. It serves as a unified data storage center, which provides a reliable data source for subsequent model training and analysis. This architectural design enhances the flexibility of data management while laying a solid foundation for the NWDAF to realize AI-native capabilities.

2.2 Model Training and Management

MTM is a core component of the NWDAF and plays a pivotal role in realizing the intelligent capabilities of the CN. The MTM architecture, as shown in Fig. 3, encompasses three key components: the model training logical function (MTLF), analytics logical function (AnLF), and ADRF. MTLF serves as the central component for training ML models within the NWDAF. It acquires the required data from the ADRF database through the data interface, where the data is preprocessed to ensure data quality, format, and consistency. Subsequently, the selected model is trained using the prepared data. Once training is completed, the trained model weights can be saved and stored in the specific model repository of the ADRF. Additionally, historical data from the ADRF can be used for external model training, with the corresponding model weights saved in the same model repository of the ADRF. The AnLF is responsible for analyzing real-time data generated by the network, using models trained by the MTLF and generating relevant analysis results. By analyzing diverse network datasets, the AnLF supports applications such as traffic prediction, network load balancing, and anomaly detection, thereby facilitating intelligent network optimization. The AnLF incorporates both data interfaces and model interfaces. The data interfaces enable the AnLF to acquire necessary network data from the ADRF and other NFs, ensuring that analyses are based on the latest data. Meanwhile, the model interfaces allow the AnLF to dynamically select and apply the most appropriate trained models, ensuring optimal performance in different scenarios. Upon completion of the model analysis, the AnLF feeds the results back to the NWDAF consumer. These results can be directly utilized for decision-making to enhance overall network intelligence. Beyond model training and analysis, model storage emerges as a critical consideration in the NWDAF. In this system, the MTLF and AnLF specifically focus on model training and analysis, without incorporating model storage. Consequently, an independent machine learning model storage module is implemented. Following model training in the MTLF, information such as version numbers and training time is stored alongside the model repository. This system facilitates version management, access control, and state monitoring through the model management function.

The design of the MTLF offers significant advantages in improving the performance and resource utilization of 6G networks, promoting AI-native capabilities in 6G. Through the coordinated operation of its three modules, the system can address single optimization objectives such as load balancing, as well as multiple optimization objectives like improving energy efficiency. Due to the complexity of network environments, a single model often cannot perform well in all scenarios. Therefore, the MTLF trains multiple models based on different NFs and scenarios, which are stored in the ADRF database. Once the NWDAF receives a service request from an NF consumer, the system deploys the most optimal model for that NF. For example, in the case of UPF scaling, the system queries the model repository for models that can meet the UPF scaling re-

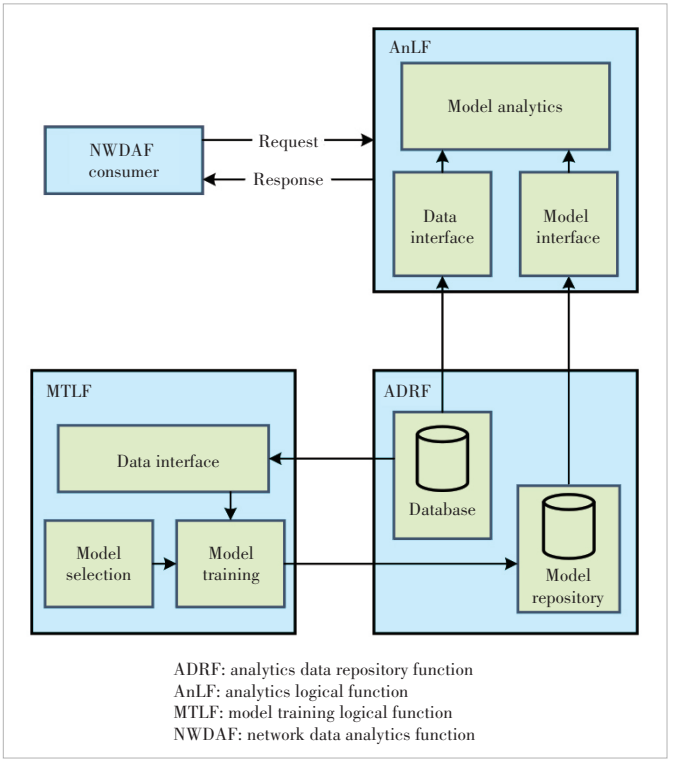


Figure 3. Model training and management

quirements. Then, based on a predefined configuration table, the optimal model is selected by considering factors such as prediction time and model deployment time. This approach improves network performance and resource efficiency, offering a practical solution for realizing AI-native capabilities in 6G networks.

3 System Deployment and Experimental Validation

To evaluate the feasibility of the proposed system, a testbed is constructed based on Kubernetes. Each NF is encapsulated in a dedicated container and managed through unified orchestration and deployment using Kubernetes. Subsequently, a UPF scaling experiment is conducted on the testbed to validate the intelligent capabilities of the system.

3.1 Deployment Strategy

The testbed leverages the open-source CN framework Open5GS as its foundation, which provides essential functions including the AMF, SMF, UPF, etc. Since Open Source 5G System (Open5GS) does not natively support the NWDAF, we develop and integrate a custom NWDAF, equipping it with capabilities for data collection and model training. To support wireless signal processing, this testbed integrates OpenAir-Interface5G (OAI) and deploys Universal Software Radio Peripheral (USRP) B210 to implement a Next-Generation Node B (gNB) and UE, effectively simulating real-world 5G network scenarios. The testbed adopts a microservice architecture utilizing containerization, encapsulating each NF as an individual container. Containers with similar NFs are strategically deployed on the same node, facilitating rapid updates, deployments, and scaling of NFs while ensuring fault isolation and improving system resilience. As illustrated in Fig. 4, the system's deployment structure comprises two servers designated as Control Plane and User Plane nodes within the CN. The Control Plane node hosts the AMF and SMF, while the User Plane node houses the UPF. Additionally, the gNB is encapsulated as a container and deployed on a dedicated Kubernetes node. To enable real communication, USRP B210 is connected to the host running the gNB container. Similarly, UE is deployed directly on its host and connected to another USRP B210 device to facilitate connectivity to the gNB.

3.2 Experiments

The experiments aim to validate the auto-scaling capabilities of the UPF and are conducted on the previously described testbed. The setup involves connecting four PCs to USRP B210 devices, utilizing OAI to simulate two user devices and two gNBs, which establish connectivity to the CN. This configuration enables two user devices to access the CN, with load generation on the UPF being achieved through data packet transmission between the user devices using the iPerf tool. The experiment simulates and verifies the process of vertical scaling. As shown in Fig. 5, vertical scaling involves increasing the resource allocation of a single pod, including CPU cores, memory capacity, and storage, to handle increased load. During the experiment, real-time data are collected and analyzed to ensure the system operates normally. Several key metrics from the AMF, SMF, and UPF are tracked to monitor the experiment process, including UPF metics such as traffic, CPU usage, and Radio Access Network (RAN) data (e.g., UE uplink/downlink traffic and bitrates). A total of 26 metrics related to UPF scaling are collected, including UPF CPU usage, the data transfer rate, and the uplink/downlink throughput of UE. These data, comprising 600 samples, are continuously collected over a 10-minute period and utilized in model training. The UPF CPU usage after 30 s is used as the prediction target, while other collected related data serve as the training param-

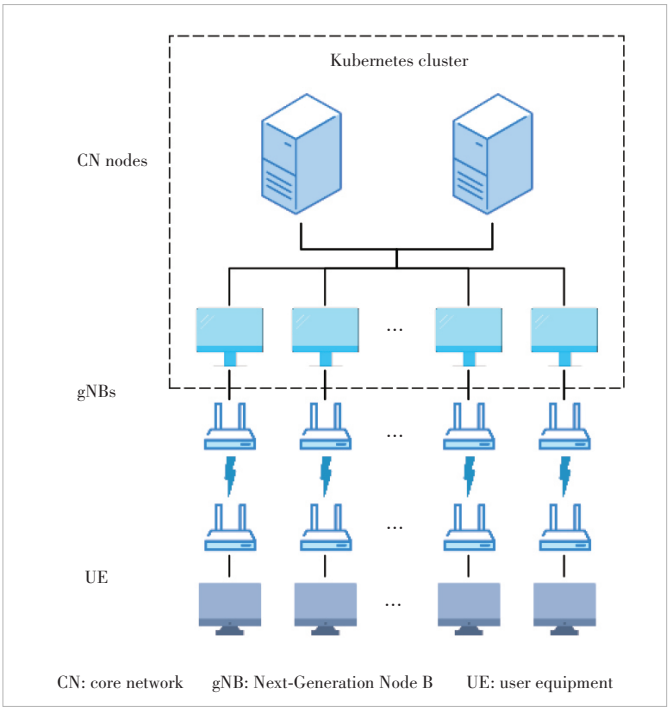


Figure 4. System deployment structure

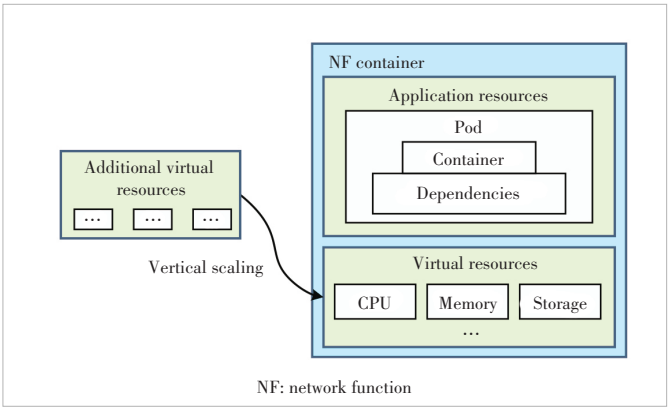


Figure 5. Vertical scaling of NF container

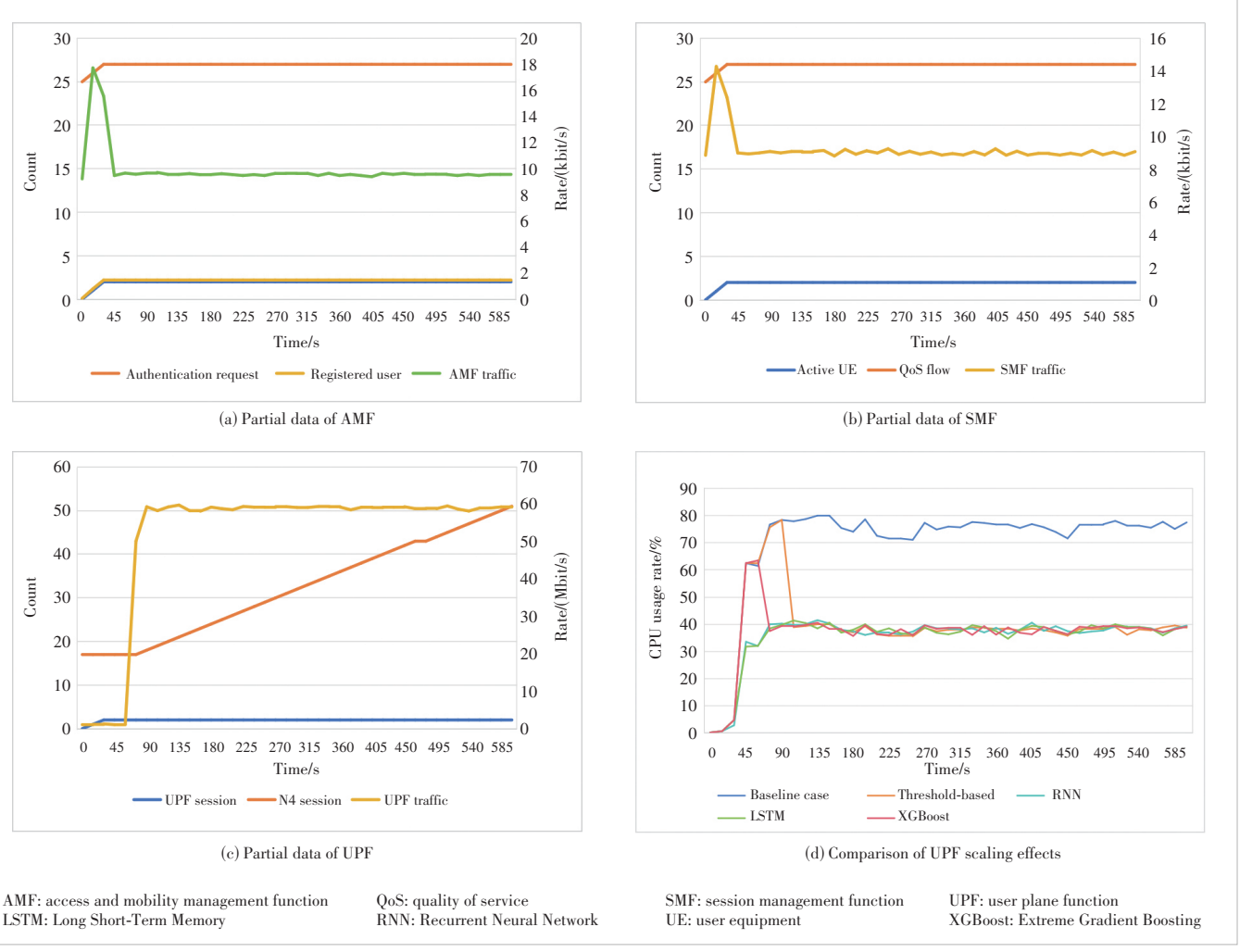
eters. When the predicted UPF CPU usage after 30 s exceeds 70%, a UPF vertical scaling operation is triggered.

To evaluate the performance of UPF scaling across various scenarios, multiple ML models are trained using the collected data. These models include LSTM[20], Extreme Gradient Boosting (XGBoost)[21], and Recurrent Neural Networks (RNNs)[22]. The trained model weights are saved as files, offering advantages in terms of shareability and reduced storage requirements. They are subsequently stored in a dedicated model repository. Finally, these models are invoked in the simulated environment, and scaling strategies are implemented based on the prediction results through the Kubernetes API.

3.3 Results

At the outset of the experiment, the relevant data from the AMF, SMF, and UPF are collected. As shown in Figs. 6a, 6b, and 6c, the data, including the number of user devices successfully connected and the number of UPF sessions, indicate that both user devices are successfully connected to the CN. Throughout the UE connection process to the CN, a large number of requests are generated, leading to an increase in network traffic of the AMF and SMF around the 15th second, followed by a quick decline. N4 sessions refer to session instances created between the SMF and UPF to ensure communication between the user plane and the control plane. Due to continuous data transmission, the number of N4 sessions increases to 50. Additionally, the UPF traffic values consistently remain around 60 Mbit/s during the communication process between the user devices, demonstrating successful data transmission between the terminals. These results indicate that when multiple user devices connect to the CN and engage in data transmission, the system successfully executes real-time data collection and analyzes the current system states.

Fig. 6d demonstrates how intensive data transmission between user devices leads to UPF scaling . The blue line represents the baseline case without scaling, while the other four lines show the results of using a threshold-based method and three ML models (LSTM, XGBoost, and RNN) for UPF scal-



(a) Partial data of AMF

(b) Partial data of SMF

(c) Partial data of UPF

(d) Comparison of UPF scaling effects

AMF: access and mobility management function
LSTM: Long Short-Term Memory
QoS: quality of service
RNN: Recurrent Neural Network
SMF: session management function
UE: user equipment
UPF: user plane function
XGBoost: Extreme Gradient Boosting

Figure 6. Data collection and algorithm comparison during the scaling process

ing. The results show that all three ML models successfully predict the increase in CPU utilization and trigger scaling. The threshold-based method also triggers scaling successfully. However, compared to the threshold-based method, all three ML models require significantly less time. In addition, Table 1 demonstrates the prediction accuracy of the three ML models. The results show that all three models achieve high accuracy, further confirming their satisfactory predictive capabilities. This experiment successfully validates the process of data collection and model analysis, demonstrating the advantages of using machine learning models for UPF scaling within the proposed architecture. Consequently, it further confirms the feasibility of the architecture presented in this paper.

Table 1. Prediction accuracy of the three models

Model	Accuracy/%
RNN	94.74
LSTM	94.87
XGBoost	86.84

LSTM: Long Short-Term Memory
RNN: Recurrent Neural Network
XGBoost: Extreme Gradient Boosting

4 Conclusions

This paper presents a novel NWDAF architecture designed to enhance the performance and scalability of AI-native 6G networks. The proposed architecture integrates data collection, model training, and analysis, providing a robust solution for dynamic network management. In addition, a testbed is established to conduct data collection and model training, validating the proposed framework with a specific focus on UPF overload and scaling scenarios. The experimental results demonstrate that the implementation of the ML models can effectively reduce the delay in UPF scaling, enhancing system responsiveness and performance. Future research directions include advancing the management of the ML models and exploring the comparative advantages of different models across various scenarios. Additionally, the integration of AI-native networks with multiple domains, such as edge computing, will be investigated to enhance real-time decision-making and resource allocation. Through these multi-domain integrations, the scalability, interoperability, and autonomous management of the network can be further enhanced.

References

[1] YANG B LIANG X, LIU S N, et al. Intelligent 6G wireless network with multi-dimensional information perception [J]. ZTE communications, 2023, 21(2): 3 – 10. DOI: 10.12142/ZTECOM.202302002
[2] CHEN Z R, ZHANG Z Y, YANG Z H. Big AI models for 6G wireless networks: opportunities, challenges, and research directions [J]. IEEE wireless communications, 2024, 31(5): 164 – 172. DOI: 10.1109/MWC.015.2300404
[3] HE S W, DONG H L, PENG S L, et al. Challenges and methods of constructing a verification system for endogenous intelligent communication in wireless networks [J]. Mobile communications, 2024, 48(7): 2 – 14. DOI: 10.3969/j.issn.1006-1010.20240629-0001
[4] GAO Y CHEN J J, LI D P. Intelligence driven wireless networks in B5G and 6G era: a survey [J]. ZTE communications, 2024, 22(3): 99 – 105. DOI: 10.12142/ZTECOM.202403012
[5] AGIWAL M, ROY A, SAXENA N. Next generation 5G wireless networks: a comprehensive survey [J]. IEEE communications surveys and tutorials, 2016, 18(3): 1617 – 1655. DOI: 10.1109/COMST.2016.2532458
[6] ZAIDI Z, FRIDERIKOS V, YOUSAF Z, et al. Will SDN be part of 5G? [J]. IEEE communications surveys and tutorials, 20(4): 3220 – 3258. DOI: 10.1109/comst.2018.2836315
[7] LUONG D H, THIEU H T, OUTTAGARTS A, et al. Cloudification and autoscaling orchestration for container-based mobile networks toward 5G: experimentation, challenges and perspectives [C]//Proc. IEEE 87th Vehicular Technology Conference (VTC Spring). IEEE, 2018: 1 – 7. DOI: 10.1109/VTCSpring.2018.8417602
[8] QUAN Q. Intelligent and autonomous management in cloud-native future networks: a survey on related standards from an architectural perspective [J]. Future Internet, 2021, 13(2), 42. DOI: 10.3390/fi13020042
[9] HE S W. An endogenous intelligent architecture for wireless communication networks [J]. Wireless networks, 2024, 30(2): 1069 – 1084. DOI: 10.1007/s11276-023-03545-9
[10] 3GPP. Technical specification group services and system aspects; achitecture enhancements for 5G system (5GS) to support network data analytics services: TS 23.288 [S]. 2023
[11] LEE J, SOLAT F, KIM T Y, et al. Federated learning-empowered mobile network management for 5G and beyond networks: from access to core [J]. IEEE communications surveys and tutorials, 2024, 26(3): 2176 – 2212. DOI: 10.1109/COMST.2024.3352910
[12] SAAD W, BENNIS M, CHEN M Z. A vision of 6G wireless systems: applications, trends, technologies, and open research problems [J]. IEEE network, 2020, 34(3): 134 – 142. DOI: 10.1109/MNET.001.1900287
[13] GKONIS P K, NOMIKOS N, TRAKADAS P, et al. Leveraging network data analytics function and machine learning for data collection, resource optimization, security and privacy in 6G networks [J]. IEEE access, 2024, 12: 21320 – 21336. DOI: 10.1109/ACCESS.2024.3359992
[14] WU W, ZHOU C H, LI M S, et al. AI-native network slicing for 6G networks [J]. IEEE wireless communications, 29(1): 96 – 103. DOI: 10.1109/mwc.001.2100338
[15] MEKRACHE A, BOUTIBA K, KSENTINI A. Combining network data analytics function and machine learning for abnormal traffic detection in beyond 5G [C]//Proc. IEEE Global Communications Conference. IEEE, 2023: 1204 – 1209. DOI: 10.1109/GLOBECOM54140.2023.10436766
[16] NISHA L K, KUMAR R. A smart data analytics system generating for 5G N/W system via ML based algorithms for the better communications [C]//Proc. 1st International Conference on Innovative Sustainable Technologies for Energy, Mechatronics, and Smart Systems (ISTEMS). IEEE, 2024: 1 – 6. DOI: 10.1109/ISTEMS60181.2024.10560068
[17] SEVGICAN S, TURAN M, GÖKARSLAN K, et al. Intelligent network data analytics function in 5G cellular networks using machine learning [J]. Journal of communications and networks, 2020, 22(3): 269 – 280. DOI: 10.1109/JCN.2020.000019
[18] MANIAS D M, CHOUMAN A, SHAMI A. An NWDAF approach to 5G core network signaling traffic: analysis and characterization [C]//Proc. IEEE Global Communications Conference. IEEE, 2022: 6001 – 6006. DOI: 10.1109/GLOBECOM48099.2022.10000989
[19] ZHANG C J, SHAN G Y, ROH B H. Fair federated learning for multitask 6G NWDAF network anomaly detection [EB/OL]. (2024-09-25) [2024-10-09]. https://ieeexplore.ieee.org/document/10693935
[20] SANTOS G L, ENDO P T, SADOK D, et al. When 5G meets deep learning: a systematic review [J]. Algorithms, 2020, 13(9): 208. DOI: 10.3390/

a13090208
[21] TEZERGIL B, ONUR E. Wireless backhaul in 5G and beyond: issues, challenges and opportunities [J]. IEEE communications surveys and tutorials, 2022, 24(4): 2579 – 2632. DOI: 10.1109/COMST.2022.3203578
[22] LY A, YAO Y D. A review of deep learning in 5G research: channel coding, massive MIMO, multiple access, resource allocation, and network security [J]. IEEE open journal of the communications society, 2021, 2: 396 – 408. DOI: 10.1109/OJCOMS.2021.3058353

Biographies

HE Shiwen (shiwen.he.hn@csu.edu.cn) is a professor at the School of Computer Science and Engineering, Central South University, China. His research interests include basic theoretical research and standard protocol development in wireless cellular/satellite/WLAN communication and networking, distributed learning and optimization computing, data mining and intelligent analysis, as well as research and development of low-level implementation theory and application technology for open programmable AI-native communication prototype systems.

PENG Shilin received his BS degree in IoT engineering from the School of Internet of Things Engineering, Hohai University, China in 2023. He is currently pursuing his MS degree in computer technology at Central South University, China. His research interest is AI-Native wireless communication.

DONG Haolei received his MS degree in computer science from the School of Computer Science, Wuhan University, China in 2019. He is currently pursuing his PhD degree in computer science at Central South University, China. His research interests include AI-Native wireless communication, 6G core networks, and knowledge graphs.

WANG Liangpeng is a senior engineer at Purple Mountain Laboratories (PML), China, specializing in wireless communication and network technologies. His research focuses on big data analytics and AI algorithms for networks, as well as knowledge graph-driven algorithms for autonomous network operations and intelligence.

AN Zhenyu is currently a senior engineer at Purple Mountain Laboratories (PML), China. His research interests include optimization theory and ultra-reliable and low latency communications.

Device Activity Detection and Channel Estimation Using Score-Based Generative Models in Massive MIMO

TANG Chenyue[1], LI Zeshen[1], CHEN Zihan[2], Howard H. YANG[1]

(1. ZJU-UIUC Institute, Zhejiang University, Haining 314400, China;
2. Singapore University of Technology and Design, Singapore 487372, Singapore)

DOI: 10.12142/ZTECOM.202501007

https://kns.cnki.net/kcms/detail/34.1294.TN.20250312.1558.002.html, published online March 13, 2025

Manuscript received: 2025-01-02

Abstract: The growing demand for wireless connectivity has made massive multiple-input multiple-output (MIMO) a cornerstone of modern communication systems. To optimize network performance and resource allocation, an efficient and robust approach is joint device activity detection and channel estimation. In this paper, we present an approach utilizing score-based generative models to address the underdetermined nature of channel estimation, which is data-driven and well-suited for the complex and dynamic environment of massive MIMO systems. Our experimental results, based on a comprehensive dataset generated through Monte-Carlo sampling, demonstrate the high precision of our channel estimation approach, with errors reduced to as low as −45 dB, and exceptional accuracy in detecting active devices.

Keywords: activity detection; channel estimation; inverse problem; score-based generative model; massive MIMO

Citation (Format 1): TANG C Y, LI Z S, CHEN Z H, et al. Device activity detection and channel estimation using score-based generative models in massive MIMO [J]. ZTE Communications, 2025, 23(1): 53 – 62. DOI: 10.12142/ZTECOM.202501007

Citation (Format 2): C. Y. Tang, Z. S. Li, Z. H. Chen, et al., "Device activity detection and channel estimation using score-based generative models in massive MIMO," *ZTE Communications*, vol. 23, no. 1, pp. 53 – 62, Mar. 2025. doi: 10.12142/ZTECOM.202501007.

1 Introduction

1.1 Motivation

The advent of the Internet of Things (IoT) era is marked by a significant increase in the number of connected devices, each capable of sensing and communicating, which has brought about a new set of challenges in network connectivity[1 – 2]. The IoT, with its expected massive device connectivity, is poised to revolutionize various aspects of daily life and socio-economic activities, from smart homes and cities to healthcare applications. These applications require ubiquitous connectivity, making massive machine-type communications (mMTC) a critical component of the upcoming 6G networks[3]. MMTC aims to provide wireless connectivity to a vast number of devices with low-complexity and low-power, which is essential for realizing IoT-based applications but also poses significant challenges in terms of network management and efficiency[4 – 6].

One of the key enablers for mMTC is the massive multiple-input multiple-output (MIMO) technology[7], which is expected to significantly improve spectral and energy efficiency at the base station (BS) level. However, a major challenge lies in acquiring accurate channel state information (CSI) for mMTC, as the pilot-aided training overhead for uplink channel estimation scales with the number of devices, which can be extremely large in a massive connection scenario[8]. A typical characteristic of mMTC traffic is its sporadic pattern, with most devices designed to remain in sleep mode for energy conservation and only a limited number active for data transmission at any given time interval[9]. This sporadic nature entails the design of joint device activity detection and channel estimation to reduce the training overhead for channel estimation.

Traditional methods for channel estimation often use dimension reduction techniques (e.g., the discrete Fourier transform) to reduce the pilot sequence length and computational complexity, which may lead to performance degradation due to the off-grid effect and energy leakage[10]. These methods also fail to capitalize on the common sparsity across different frequency bands. To address these limitations, a novel sparse Bayesian learning (SBL) framework for joint device activity detection and channel estimation has been proposed, exploiting additional sparsity structures to significantly enhance sparse

TANG Chenyue and LI Zeshen are co-first authors.

recovery performance[11].

Inspired by the potential of score-based generative models in specialized applications such as magnetic resonance imaging (MRI) reconstruction[12-13], we introduce a training and inference algorithm for wireless channel estimation using score-based generative models in massive MIMO communication scenarios. This approach models the log-distribution of channels by learning the high-dimensional gradient, known as the score, providing a distribution learning framework for modeling high-dimensional millimeter-wave (mmWave) MIMO channels in a stochastic environment. Unlike traditional methods, our approach uses score-based generative models to learn the score of the distribution in an unsupervised manner, independent of pilot symbols. Device activity detection and probabilistic channel estimation are achieved by sampling from the posterior distribution using annealed Langevin dynamics, tackling challenges in out-of-distribution settings, wide signal-to-noise ratio (SNR) ranges, and interference scenarios.

1.2 Related Work

Many problems in engineering applications, ranging from signal processing and computer vision to machine learning and statistics, can be formulated as linear inverse problems[14]. To solve these linear inverse problems, researchers have proposed various approaches, such as compressed sensing methods[15] and deep learning techniques[16]. Pre-trained generative priors have also been used in solving linear inverse problems, surpassing classical compressed sensing approaches[17-18]. With the emergence of deep generative models in density estimation[19-21], there has been a surge of interest in developing linear inverse algorithms with data-driven priors[22-23]. Owing to the powerful representational capabilities of deep generative models, they can effectively learn accurate prior knowledge given sufficient data samples[14]. Their potential in solving linear inverse problems is gaining increasing attention.

Massive connectivity is a key requirement for future wireless cellular networks to support mMTC[5]. In large-scale wireless cellular networks, user detection and channel estimation can be viewed as high-dimensional linear inverse problems, as scheduling a large number of occasionally active users on a separate control channel may incur significant overhead. Studies such as Refs. [24] and [25] investigate a random access protocol in which each active user picks one of the orthogonal signature sequences at random and sends it to the BS, and a connection is established if the selected preamble is not used by the other users. Refs. [5] and [26] propose the use of the approximate message passing (AMP)[27] algorithm for joint user activity detection and channel estimation, and further show that a state evolution analysis[28] of the AMP algorithm allows an analytic characterization of the missed detection and false alarm probabilities for device detection.

As a novel class of generative models, diffusion models (DM), also known as score-based generative models, have achieved remarkable performance in density estimation and image generation[20, 29]. Originally, DM was introduced for unconditional image generation; however, they have since been widely applied to conditional probability distributions, enabling tasks such as conditional image generation[30]. Supervised end-to-end training of deep learning-based methods has been successfully applied to wireless MIMO channel estimation[31-32], introducing a powerful and robust deep learning algorithm in the form of the learned denoising approximate message passing (L-DAMP) algorithm[33]. Furthermore, Ref. [34] employs annealed Langevin dynamics and score-based models to efficiently train generative models on simulated datasets, achieving performance superior to that of generative adversarial networks (GANs).

1.3 Contributions

The principal contributions of this paper are summarized as follows:

• We introduce an approach that leverages score-based generative models to achieve joint active device detection and channel estimation for massive MIMO communications. Our solution delivers accurate estimates without imposing any assumptions on the dimensionality or sparsity of the channels, thereby providing a flexible and robust method for real-world applications.

• We generate simulated data with varying sizes and complexity, closely capturing the diverse and dynamic nature of massive MIMO environments. This capability allows our model to be trained and tested under conditions that accurately reflect real-world wireless propagation scenarios.

• Through extensive numerical simulations, we validate the effectiveness of our method. The results indicate that the accuracy of active device detection exceeds 98% under high SNR conditions. Additionally, the normalized mean square error (NMSE) can be reduced to as low as −45 dB, highlighting the superior performance of our approach in channel state estimation and active user detection in massive MIMO systems.

1.4 Organization

The remainder of this paper is organized as follows. Section 2 presents the massive MIMO system model, the inverse problem, and the procedures involved in score-based generative models. Section 3 details the training phase of the proposed method, focusing on the generation of channel data and the training of the score function, as well as the inference (testing) stage. Section 4 provides numerical simulation results and discussions. Sections 5 and 6 conclude the paper and present the future work.

2 Preliminaries

2.1 Massive MIMO

Massive MIMO is a key technology for next-generation wireless communication systems, characterized by the deployment

of a large number of antennas at the BS to serve multiple users simultaneously[35]. This configuration allows for significant improvements in spectral efficiency, energy efficiency, and overall system performance. By leveraging spatial multiplexing and beamforming techniques, massive MIMO can effectively mitigate interference, increase data rates, and improve the reliability of wireless links. The large number of antennas enables the BS to exploit the spatial diversity of the channel, leading to more precise CSI estimation and better resource allocation. As a result, massive MIMO serves as a key enabler for next-generation wireless networks, including 5G and beyond, addressing the growing demand for high-speed, low-latency, and high-capacity communication services[36].

2.2 Inverse Problem

Inverse problems are ubiquitous in various scientific and engineering fields, where the goal is to infer the parameters or states of a system from observed data. A common linear model used to describe inverse problems is expressed as:

$$\boldsymbol{Y} = \boldsymbol{XP} + \boldsymbol{Z} \tag{1},$$

where $\boldsymbol{Y}$ represents the observed data, $\boldsymbol{X}$ is the unknown system matrix or operator, $\boldsymbol{P}$ is the known parameter vector, and $\boldsymbol{Z}$ is the noise term. The objective is to estimate $\boldsymbol{X}$ from $\boldsymbol{Y}$ and $\boldsymbol{P}$, but reconstructing the underlying causes from their observed effects is inherently complex, particularly in real-world scenarios. This challenge is further compounded by the presence of noise and the potential for the problem to be ill-posed, meaning that solutions may not exist, may not be unique, or may be excessively sensitive to noise. To address these issues, regularization techniques such as Tikhonov or total variation regularization are commonly applied. These methods add constraints to the problem to stabilize the solution[37].

In situations where the number of unknowns surpasses the number of measurements, referred to as under-determined problems, the challenge intensifies. The disparity between the number of unknowns and the available data leads to a scenario with an infinite number of potential solutions to $\boldsymbol{X}$ that could align with the equation $\boldsymbol{Y} = \boldsymbol{XP} + \boldsymbol{Z}$. This scenario is particularly problematic as it significantly increases the risk of inaccurate or unstable solutions[13]. To combat these difficulties, optimization methods and Bayesian approaches are employed. These strategies incorporate prior knowledge and provide a framework for managing the uncertainty associated with the estimates. Furthermore, recent progress in machine learning and generative modeling has introduced innovative approaches to address these challenges. These advancements offer new methods to handle the instability and uncertainty inherent in under-determined problems, thereby improving the reliability and accuracy of the solutions derived from noisy and incomplete data.

Specifically, score-based generative models have shown promise in addressing under-determined inverse problems by leveraging the underlying data distribution to generate plausible solutions. These models represent a powerful approach capable of capturing the complex structures of high-dimensional data distributions without explicit parametric forms, making them especially suitable for applications where data distribution is complex or not easily characterized by traditional models[17]. Such applications are particularly relevant for real-world wireless environments.

2.3 Score-Based Generative Model

A core hypothesis of this study is that the characteristics of wireless channels can be represented as samples drawn from a common probability distribution, which has been widely adopted in both theoretical and practical wireless communications research[38]. Score-based generative models, which have demonstrated their effectiveness on natural image benchmark datasets, are a class of generative models that generate data by estimating the gradient of the data distribution[13]. This approach diverges from traditional generative modeling techniques, which often rely on explicit parameterization of the data distribution. Instead, score-based generative models learn the gradient field of the data distribution in a non-parametric fashion, providing a flexible framework for capturing complex data distributions[20]. Fig. 1 fully displays the process of handling inverse problems using a score-based generative model.

2.3.1 Learning Score Function

The score function for a point $\boldsymbol{X}$ is represented as:

$$\boldsymbol{\psi}_{\boldsymbol{X}}(\boldsymbol{X}) = \nabla \log p_{\boldsymbol{X}}(\boldsymbol{X}) \tag{2},$$

where $\boldsymbol{X}$ denotes the data point, $p_{\boldsymbol{X}}(\boldsymbol{X})$ is the probability density distribution of this data point, and $\boldsymbol{\psi}_{\boldsymbol{X}}(\boldsymbol{X})$ is a matrix of size $M \times N$. The score function encapsulates the local density

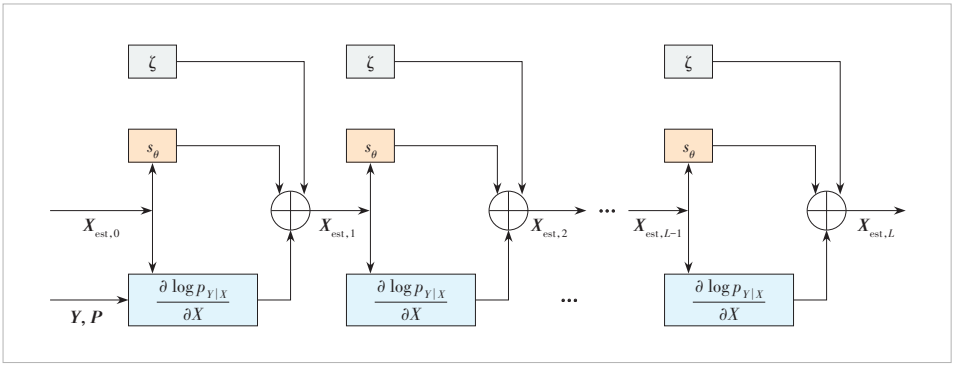


Figure 1. A step-by-step process for estimating X by employing a score-based model in conjunction with the known matrices Y and P

information of the data distribution, which is instrumental in the generative process. In practice, $\boldsymbol{\psi}_{\boldsymbol{X}}(\boldsymbol{X})$ can be used to guide the optimization process for channel estimation by iteratively updating the channel estimate in the direction that maximizes the likelihood of the observed data. For example, if the score function indicates a high likelihood of a certain channel coefficient being non-zero, the algorithm can focus on refining the estimate of that coefficient, leading to more accurate channel estimation overall. The goal is to learn a model s_θ capable of generating $s_\theta(\boldsymbol{X})$ to approximate $\boldsymbol{\psi}_{\boldsymbol{X}}(\boldsymbol{X})$.

2.3.2 Denoising Score Matching

While $\boldsymbol{\psi}_{\boldsymbol{X}}$ and the explicit score matching $\mathbb{E}_{\boldsymbol{X}\sim p_{\boldsymbol{X}}}\left[\left\| s_\theta(\boldsymbol{X}) - \boldsymbol{\psi}_{\boldsymbol{X}}(\boldsymbol{X})\right\|_2^2\right]$ are often intractable, denoising score matching is proposed to address this issue. Ref. [39] demonstrates that the loss function $\mathcal{L}(s_\theta)$ we used for training is equivalent to the loss function of the explicit score matching, as long as $\log p_{\tilde{\boldsymbol{X}}|\boldsymbol{X}}(\tilde{\boldsymbol{X}}|\boldsymbol{X})$ is differentiable with respect to $\tilde{\boldsymbol{X}}$. This approach transforms the task of learning the score function of the original data distribution (which is nearly impossible in the real world) into learning the score of the perturbed distribution by using $\mathcal{L}(s_\theta)$. By synthesizing corrupted data samples $\tilde{\boldsymbol{X}}$ and learning the score of the conditional distribution $p_{\tilde{\boldsymbol{X}}|\boldsymbol{X}}$, the following objective is used:

$$\mathcal{L}(s_\theta) = \mathbb{E}_{\boldsymbol{X}\sim p_{\boldsymbol{X}},\tilde{\boldsymbol{X}}\sim p_{\tilde{\boldsymbol{X}}}}\left[\left\| s_\theta(\tilde{\boldsymbol{X}}) - \nabla \log p_{\tilde{\boldsymbol{X}}|\boldsymbol{X}}(\tilde{\boldsymbol{X}}|\boldsymbol{X})\right\|_2^2\right] \tag{3}$$

Since IoT allows the use of arbitrary noise distributions for training and learning the score at arbitrarily perturbed inputs, we set the perturbation $\boldsymbol{U}$ as i.i.d. Gaussian, with zero mean and covariance matrix $\sigma_U^2\boldsymbol{I}$, i.e.,

$$\nabla \log p_{\tilde{\boldsymbol{X}}|\boldsymbol{X}}(\tilde{\boldsymbol{X}}|\boldsymbol{X}) = -\boldsymbol{U}/\sigma_U^2 \tag{4}$$

A learnable model proposed by Ref. [20] is used to learn s_θ. The model (in our work, such a deep neural network described in Section 3) uses a weighted version of $\mathcal{L}(s_\theta)$ at multiple noise levels to train a single score-based model for an individual datum within a batch, represented by:

$$\mathcal{L}_{\text{score}}(\theta) = \mathbb{E}_{j,\boldsymbol{X}\sim p_{\boldsymbol{X}},\boldsymbol{U}_j\sim p_{\boldsymbol{U}_j}}\left[\sigma_{U_j}^2\left\| s_\theta(\boldsymbol{X}+\boldsymbol{U}_j) + \frac{\boldsymbol{U}_j}{\sigma_{U_j}^2}\right\|_2^2\right] \tag{5}$$

Weighing the predicted score at each noise level is to formulate denoising score matching as a variance-exploding (VE) diffusion process[40].

2.3.3 Posterior Sampling Using Score Functions

Once the score function is learned, it can be used to perform posterior sampling, which is a key step in the channel estimation process. Posterior sampling involves drawing samples from the posterior distribution of the CSI conditioned on the received pilot symbols. Given the known matrices $\boldsymbol{Y}$ and $\boldsymbol{P}$, the posterior distribution of matrix $\boldsymbol{X}$ can be expressed using the Bayes' rule:

$$p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{X}|\boldsymbol{Y}) = \frac{p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{Y}|\boldsymbol{X})\cdot p_{\boldsymbol{X}}(\boldsymbol{X})}{p_{\boldsymbol{Y}}(\boldsymbol{Y})} \tag{6}$$

Expanding the logarithm of the posterior distribution, we get

$$\log p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{X}|\boldsymbol{Y}) = \log p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{Y}|\boldsymbol{X}) + \log p_{\boldsymbol{X}}(\boldsymbol{X}) - \log p_{\boldsymbol{Y}}(\boldsymbol{Y}) \tag{7}$$

Taking the gradient with respect to $\boldsymbol{X}$, we obtain

$$\nabla \log p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{X}|\boldsymbol{Y}) = \nabla \log p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{Y}|\boldsymbol{X}) + \nabla \log p_{\boldsymbol{X}}(\boldsymbol{X}) \tag{8}$$

since $\nabla \log p_{\boldsymbol{Y}}(\boldsymbol{Y}) = 0$. For all $\boldsymbol{Y}$, the gradient of the posterior distribution simplifies to:

$$\boldsymbol{\psi}_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{X}|\boldsymbol{Y}) = \boldsymbol{\psi}_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{Y}|\boldsymbol{X}) + \boldsymbol{\psi}_{\boldsymbol{X}}(\boldsymbol{X}) \tag{9}$$

This result shows that the gradient of the posterior distribution is a combination of the gradient of the likelihood function and the gradient of the prior distribution. The likelihood function is derived from $\boldsymbol{Y}$, while the prior distribution is learned using the score-based generative model.

2.3.4 Annealed Langevin Dynamics for Posterior Sampling

To sample from the posterior distribution, we use annealed Langevin dynamics, which is an iterative process that updates the channel estimate $\boldsymbol{X}_{\text{est}}$ in a manner that maximizes the posterior probability. We introduce time-varying hyperparameters α_t and β_t as an enhancement, based on the method proposed in Ref. [41]. The update rule for annealed Langevin dynamics is given by

$$\boldsymbol{X}_{\text{est},i+1} = \boldsymbol{X}_{\text{est},i} + \alpha_i \cdot \left(\nabla \log p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{Y}|\boldsymbol{X}_{\text{est},i}) + \nabla \log p_{\boldsymbol{X}}(\boldsymbol{X}_{\text{est},i})\right) + \sqrt{2\beta\cdot\alpha_i}\cdot\sigma_{U_i}\cdot\zeta \tag{10}$$

where α_i is the step size that decays over time; β is a hyperparameter that controls the amount of noise added to the update; σ_{U_i} is the noise level at the i-th step; $\zeta\sim\mathcal{CN}(0,\boldsymbol{I})$ is Gaussian noise added to maintain diversity in the samples. The parameters α_i, β, and σ_{U_i} are critical for the performance of the proposed method. The learning rate α_i is chosen through a grid search to balance convergence speed and accuracy. The initial value of the regularization parameter β is empirically set to 0.9 for robustness to noise. The noise variance σ_{U_i} is estimated from the training data using a maximum likelihood approach.

The gradient of the likelihood function $\nabla \log p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{Y}|\boldsymbol{X}_{\text{est},i})$ can be derived from $\boldsymbol{P}$. For Gaussian noise, this gradient is given by

$$\nabla \log p_{Y|X}(\boldsymbol{Y}|\boldsymbol{X}_{\text{est},i}) = \frac{(\boldsymbol{X}_{\text{est},i}\boldsymbol{P} - \boldsymbol{Y})\boldsymbol{P}^H}{\sigma^2_{\text{pilot}}} \tag{11}.$$

The term $\nabla \log p_X(\boldsymbol{X}_{\text{est},i})$ represents the gradient of the prior distribution, which is learned using the score-based generative model. This gradient is approximated by the learned score function $s_\theta(\boldsymbol{X}_{\text{est},i})$.

3 System Model

Consider a single-cell massive MIMO network, where a BS equipped with M antennas serves N potential users, denoted by the set $\mathcal{N} = \{1,\cdots, N\}$. Each user device is equipped with a single antenna. This setup is typical for an uplink massive access scenario, where the BS efficiently manages data transfer and communication from numerous users within its coverage area. Fig. 2 illustrates an example.

In our system model for device activity detection and channel estimation, the sporadic nature of user traffic can be characterized by a user activity indicator for each user. We denote this indicator by

$$\lambda_n = \begin{cases} 1, & \text{if user } n \text{ is active} \\ 0, & \text{otherwise} \end{cases} \tag{12}.$$

The probability of a user being active is ϵ, and the probability of being inactive is $1-\epsilon$, such that $\Pr[\lambda_n = 1] = \epsilon$ and $\Pr[\lambda_n = 0] = 1-\epsilon$. The set of active users within a coherence block is defined as $\mathcal{K} = \{n:\lambda_n = 1\}$, and the number of active users is $K = |\mathcal{K}|$.

The transmitted signal for each user n is given by

$$x_n = \lambda_n h_n \tag{13},$$

where h_n represents the channel coefficient for user n. The matrix $\boldsymbol{X}$ is formed by stacking the transmitted signals of all users, i.e., $\boldsymbol{X} = [x_1,\cdots, x_N]^{\mathrm{T}}$.

During the training phase, the BS receives a matrix $\boldsymbol{Y}$, which is modeled as the product of the transmitted signal matrix $\boldsymbol{X}$, the pilot matrix $\boldsymbol{P}$, and the addition of additive white Gaussian noise $\boldsymbol{Z}$. The model can be expressed as

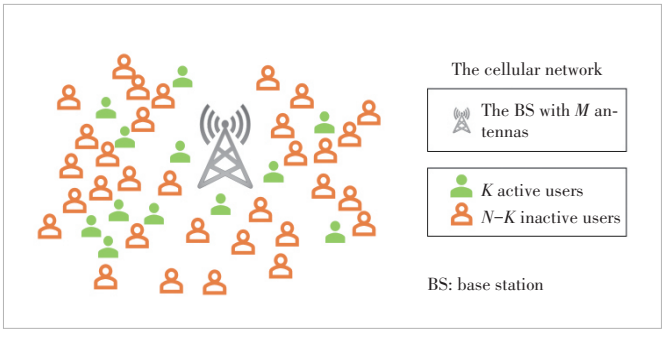


Figure 2. System model of a massive device communication network

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{P} + \boldsymbol{Z} \tag{14}.$$

Here, the channel state information matrix $\boldsymbol{X} \in \mathbb{C}^{M\times N}$, which is a complex matrix of size M times N, where M represents the number of receive antennas and N denotes the total number of users. $\boldsymbol{Z}$ is the Gaussian noise matrix with elements distributed as $\mathcal{CN}[0, \sigma^2 I]$. $\boldsymbol{P}$ is the pilot matrix where each entry is a randomly chosen (fixed for all test samples) quadrature phase shift keying (QPSK) symbol with unit amplitude and low-resolution phase. Pilot symbols L_{pilot} are selected from a pre-designed codebook, with each symbol p_i belonging to $\mathbb{C}^N$. These pilot symbols are utilized to facilitate the estimation process. The transmitted pilot matrix $\boldsymbol{P}$ is constructed from these symbols, and it is common practice in communication standards to pre-specify these pilot sequences.

The pilot density is defined by $\alpha = L_{\text{pilot}}/N$. When $\alpha < 1$, it implies that the number of received pilots is less than the total number of possible pilot transmissions, i.e., $ML_{\text{pilot}} < MN$. This situation leads to an under-determined inverse problem for channel estimation, where there are more unknowns (channel coefficients) than the number of equations provided by the received pilots.

Following the methodology outlined in Ref. [34], we employ a score-based generative model to accomplish two critical tasks: channel estimation and device activity detection. The pseudocode is shown in Algorithm 1. This model operates on a data-driven approach, which is particularly effective in addressing under-determined scenarios. In under-determined problems, the number of unknowns exceeds the number of equations, making the system's solution unstable and sensitive to noise. However, our approach can stabilize these solutions by learning the underlying data distribution, thus providing a robust framework for estimation.

Algorithm 1: Device activity detection and channel estimation via score-based generative models in massive MIMO systems

Input: Pilot matrix $\boldsymbol{P}$, received pilots $\boldsymbol{Y}$, pretrained score-based model s_θ, received noise power σ^2_{pilot}, inference noise levels $\sigma^2_{U_i}$, hyperparameters L, Q, α_0, β, and $r < 1$.

Generate random initial estimate: $\boldsymbol{X}_{\text{est},0} \sim \mathcal{CN}(0, I)$

For $i = 1, 2, \cdots, L$

Set annealed noise level $\sigma \leftarrow \sigma_{U_i}$.

For $q = 1, 2, \cdots, Q$

Generate annealing noise $\zeta \sim \mathcal{CN}(0, I)$.

$$\boldsymbol{X}_{\text{est},q} \leftarrow \boldsymbol{X}_{\text{est},q-1} + \alpha_0 \cdot r^i \cdot \frac{(\boldsymbol{X}_{\text{est}}\boldsymbol{P} - \boldsymbol{Y})\boldsymbol{P}^H}{\sigma^2_{\text{pilot}} + \sigma^2} +$$

$$\alpha_0 \cdot r^i \cdot s_\theta(\boldsymbol{X}_{\text{est}}) + \sqrt{2\beta \cdot \alpha_0 \cdot r^i} \cdot \sigma \cdot \zeta$$

Count the number of zero rows in $\boldsymbol{X}_{\text{est}}$ to find $N - K$.

Output: Estimate channel matrix $\boldsymbol{X}_{\text{est}}$, and then get the NMSE and activity detection accuracy.

The objective of our model is to estimate the CSI using the

received pilot matrix $\boldsymbol{Y}$ and the known pilot matrix $\boldsymbol{P}$, and to determine the number of inactive users indicated by $\lambda_n = 0$ in the channel matrix $\boldsymbol{X}$. The process is divided into two main phases: training and inference.

1) Training phase: This initial step involves training the score-based generative model by minimizing the loss function as detailed in Section 2.3.2. To compute the score function s_θ, we train a deep neural network, as depicted in Fig. 3. The neural network is trained to learn the score function that approximates the gradient of the log-likelihood of the data distribution. Moreover, it is fully convolutional, enabling it to process matrices of varying sizes, which is crucial for the dynamic nature of massive MIMO systems. This is a one-time setup process for the wireless device, typically conducted offline using a high-performance computing server and a dataset comprising either precise channel measurements or simulated channel data. The loss function quantifies the discrepancy between the model's predictions and the actual data, guiding the model to learn the data distribution effectively.

Particularly, our approach employs a Monte-Carlo simulation to generate synthetic massive MIMO channel data. User positioning is modeled to simulate random distribution within a defined area, reflecting real-world spatial randomness. Path loss is calculated using $\text{Path Loss/dB} = 128.1 + 37.6 \log_{10} d$ [42], which is a standard model for signal attenuation in wireless communication. Firstly, it allows for the modeling of complex channel behaviors by simulating a large number of random variables, which is essential for accurately representing the multipath fading effects in wireless communication channels[43]. Secondly, this approach facilitates the assessment of system performance under various conditions, providing a robust framework for optimizing and understanding the behavior of massive MIMO systems[36]. Then, the neural network begins with 2D downsampling and convolutional layers designed to extract meaningful features from these input datasets. To enhance the model's ability to learn complex patterns, Rectified Linear Unit (ReLU) activation functions are adopted to introduce nonlinearity. The model then employs 2D upsampling with additional convolutional layers to reconstruct the data to its original dimensions. In the closing act of the methodology, a 2D average pooling layer serves to compress feature maps, enhance noise resilience, and streamline subsequent layers by reducing dimensionality and focusing on dominant features.

2) Inference (testing) phase: In this phase, channel estimation is treated as an optimization problem and solved using the iterative algorithm presented in Sections 2.3.3 and 2.3.4. The pre-trained model, combined with the received pilots, is utilized to recover the CSI. This phase is designed to operate independently of the training stage, enabling adaptability to various real-world conditions, including interference and quantization effects on the received pilots.

The complexity of each step is analyzed as follows.

1) Initialization: The initialization step involves generating a random initial estimate $\boldsymbol{X}_{\text{est},0} \sim \mathcal{CN}(0, I)$, which has a complexity of $O(MN)$, where M and N are the dimensions of the channel matrix.

2) Outer loop $i = 1, 2, \cdots, L$: Setting annealed noise level $\sigma \leftarrow \sigma_{U_i}$ involves negligible computational complexity.

3) Inner loop $q = 1, 2, \cdots, Q$: Generating annealing noise $\zeta \sim \mathcal{CN}(0, I)$ has a complexity of $O(MN)$. The update step for $\boldsymbol{X}_{\text{est},q}$ involves several matrix operations:

- $\boldsymbol{X}_{\text{est},q} \leftarrow \boldsymbol{X}_{\text{est},q-1} + \alpha_0 \cdot r^i \cdot \frac{(\boldsymbol{X}_{\text{est}}\boldsymbol{P} - \boldsymbol{Y})\boldsymbol{P}^H}{\sigma_{\text{pilot}}^2 + \sigma^2}$: This step involves matrix multiplication and division, with a complexity of $O(MNP)$, where P is the number of pilots.
- $+\alpha_0 \cdot r^i \cdot s_\theta(\boldsymbol{X}_{\text{est}})$: The complexity of this step depends on the model s_θ, assumed to be $O(f(MN))$, where f is a function of the model complexity.
- $+\sqrt{2\beta \cdot \alpha_0 \cdot r^i} \cdot \sigma \cdot \zeta$: This step has negligible complexity.

4) Counting zero rows: Counting the number of zero rows in $\boldsymbol{X}_{\text{est}}$ to find $N - K$ has a complexity of $O(MN)$.

Thus, the total time complexity is dominated by the inner loop operations, particularly the matrix multiplications and the model s_θ evaluation. Therefore, the total complexity

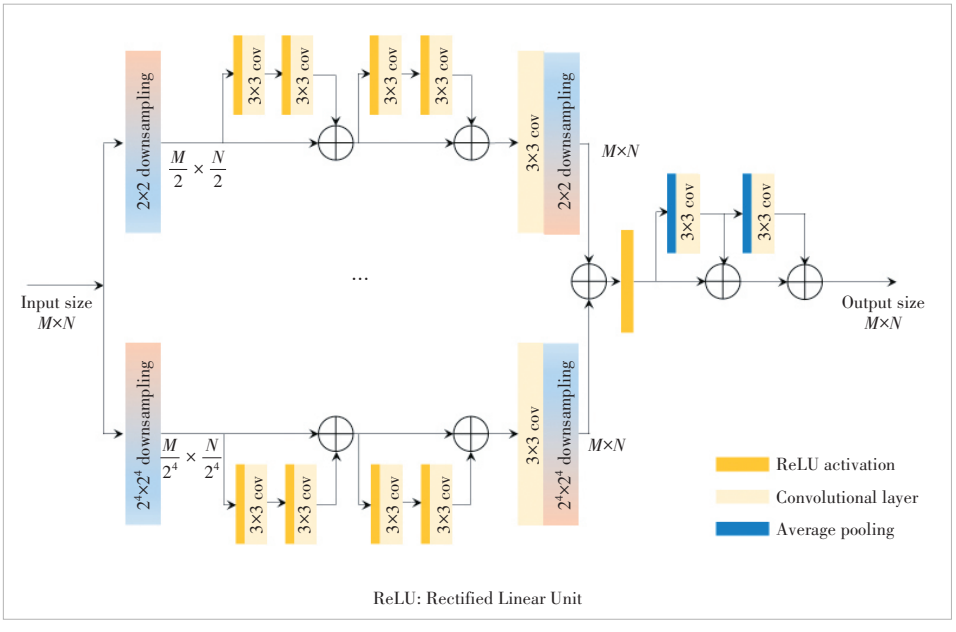


Figure 3. An elaborate schematic representation of model s_θ utilizing the RefineNet architecture. This fundamental block is cascaded D times in sequence

is approximately $O(LQ \cdot MNP + LQ \cdot f(MN))$. The space complexity is primarily determined by the storage requirements for $\boldsymbol{X}_{\text{est}}$ and other intermediate variables, which is $O(MN)$.

4 Experiments

We conduct simulations with different configurations of M (the number of antennas), N (the total number of users), and K (the number of active users) to generate input channel matrices that vary in size and complexity via Monte-Carlo methodology. This approach allows us to assess the efficacy of our model under diverse conditions, verifying its robustness and capability to accurately mirror the dynamics of realistic scenarios. Additionally, we perform a comparative analysis of our proposed score-based generative model against the traditional linear minimum mean square error (LMMSE) method for channel estimation. This comparison is conducted across various SNR levels to evaluate their performance in terms of NMSE and activity detection accuracy. All experiments are conducted using PyTorch on an NVIDIA RTX 3090 GPU.

The channel matrix is initialized by assigning random positions to users and computing the path loss as a function of their distance from the BS. Subsequently, it constructs the channel coefficients using complex Gaussian random variables to simulate the multipath fading effects and compiles these into a data matrix for each simulation iteration. After generating datasets, we assess the performance of Algorithm 1 through simulations, spanning various SNR levels. Our evaluation criteria include the accuracy of channel estimation, the error rates throughout a simulated communication system that employs coding, and the computational overhead associated with both training and inference phases. To mimic real-world deployment scenarios, we examine situations where the algorithm is challenged with data distributions that differ from those encountered during training. This assessment is conducted without any foreknowledge of the test environment's characteristics, without modifying the model to adapt to the new distribution, and without conducting any additional training specifically for the test conditions.

For fine-tuning the hyperparameters in our channel estimation methods, we utilize a subset of 500 channel realizations sampled from the training distribution. In the testing phase, we create a fresh dataset consisting of 50 channel realizations for each target distribution, ensuring that the random seed used differs from those used in the training and validation phases. For the pilot signals $\boldsymbol{P}$, we construct matrices with dimensions $N \times L_{\text{pilot}}$, filled with QPSK elements that are randomly selected to represent unit-power, two-bit phase-quantized random beamforming vectors. To standardize the channel measurements, we apply normalization using the mean channel power calculated from the training dataset, which is derived from all training samples and their respective entries. The average SNR is then determined using the formula $N/\sigma^2_{\text{pilot}}$, where N is the number of transmit antennas and σ^2_{pilot} is the variance of the pilot signals.

Our proposed model demonstrates rapid convergence, as indicated by the swift reduction in training loss during the initial steps (Fig. 4), stabilizing at a low value by the completion of training.

In our comparative analysis of channel estimation techniques, the proposed score-based generative model outperforms the LMMSE method (Fig. 5). The traditional LMMSE method[44] exhibits higher noise levels and is notably less accurate in estimating user activity rates, especially in poor channel conditions. Only under sufficiently good channel conditions can the traditional method approach the performance of our generative learning approach.

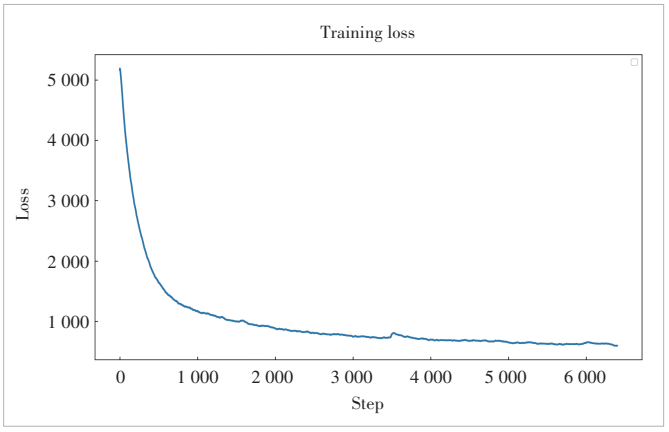


Figure 4. Training loss of the score-based generative model over steps

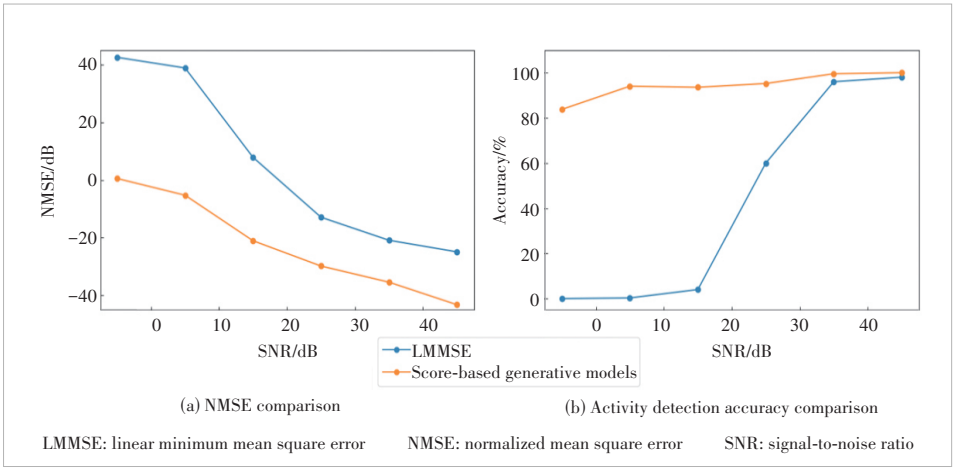


Figure 5. Performance comparison of channel estimation methods

We also input different datasets to evaluate the algorithm's performance. Three sets of comparative experiments were conducted to observe the impact of varing M, N, and K on the algorithm's effectiveness (Fig. 6). The overall experimental results indicate the following:

1) Performance improvement: Our proposed method demonstrates a substantial enhancement in performance compared to traditional methods. This is particularly evident when evaluating the impact of varying M, N, and K on the algorithm's effectiveness. Our method consistently shows lower NMSE and higher accuracy across different configurations, indicating a superior capability in channel estimation and device activity detection.

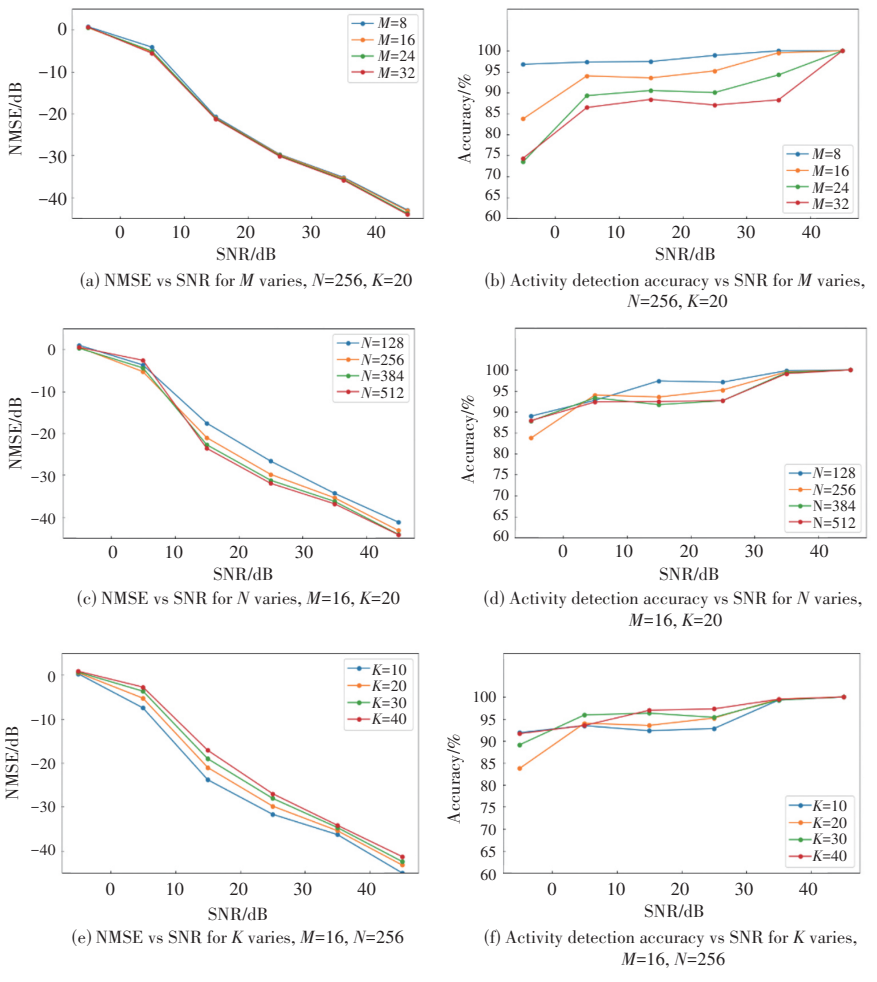
2) Robustness: The proposed score-based generative model demonstrates robustness under varying channel conditions, maintaining low estimation errors despite changes in channel conditions.

The numerical comparative analysis is as follows:

1) When only M (number of receiving antennas) varies, the maximum absolute difference in NMSE across all SNR levels is only 1.15 dB (when SNR=10 dB), indicating a minimal impact on channel estimation. However, due to the increase in matrix size, the sensitivity to data increases, and the accuracy of active detection is poor under very poor channel conditions. The smaller the M, the faster the accuracy approaches 100% (e.g., M=8, 16). Nevertheless, a perfect accuracy rate of 100% can be achieved when SNR=45 dB.

2) When only N (total number of devices) varies, the NMSE curve indicates a slight overall improvement in model performance. This suggests that our model is particularly suitable for massive MIMO scenarios. Meanwhile, the overall active detection accuracy tends towards 100% more rapidly as the SNR increases.

3) Changes in K (number of active users) do not affect the shape of the channel. The active detection accuracy remains high even under the worst channel conditions. Moreover, the CSI estimation becomes closer to the ground truth as the number of active users decreases. This is applicable to IoT scenarios, where devices are typically designed to remain inactive most of the time to conserve energy, with only a few devices active transmitting data at any given interval. This indicates that using our model to assess device activity rates and perform more accurate channel estimation could optimize device activity patterns in the future, further reducing energy consumption and improving energy efficiency.



(a) NMSE vs SNR for M varies, N=256, K=20
(b) Activity detection accuracy vs SNR for M varies, N=256, K=20
(c) NMSE vs SNR for N varies, M=16, K=20
(d) Activity detection accuracy vs SNR for N varies, M=16, K=20
(e) NMSE vs SNR for K varies, M=16, N=256
(f) Activity detection accuracy vs SNR for K varies, M=16, N=256

NMSE: normalized mean square error SNR: signal-to-noise ratio

Figure 6. Diagrams for simulaion results

5 Conclusions

In this paper, we propose a novel method for joint device activity detection and channel estimation in massive MIMO networks, enabling accurate channel estimation to enhance energy efficiency and communication performance.

We employ score-based generative models, an innovative generative approach that integrates deep neural networks without making any assumptions about the received pilot matrix, the transmitted pilot matrix, and the pilot density. During our simulation experiments, we generated a comprehensive dataset using Monte-Carlo sampling. Since the deep neural network framework used to learn the scoring function is fully convolutional, the model can flexibly adapt to inputs of various sizes. We conducted a series of com-

parative experiments under varying conditions, including varying numbers of antennas, total users, and active users. The results demonstrate that as channel conditions improve, channel estimation is highly precise, with errors reduced to as low as −45 dB, and the detection of active devices is exceptionally accurate. As the number of users increases, the NMSE decreases, indicating that our approach is highly suitable for massive MIMO scenarios. Moreover, a smaller number of active users indicates a sparser channel matrix, yet changes in activity have a minimal impact on our model's performance, confirming that our method is entirely data-driven.

6 Future Work

The proposed score-based generative model for joint device activity detection and channel estimation demonstrates significant potential for application in next-generation wireless systems. Future work will explore the adaptation of this method to mmWave channels, which have unique characteristics, such as higher frequency bands and more severe path loss. Additionally, we plan to investigate the integration of this approach into 5G and 6G deployments, where massive connectivity and high spectral efficiency are critical requirements.

References

[1] CHEN X M, NG D W K, YU W, et al. Massive access for 5G and beyond [J]. IEEE journal on selected areas in communications, 2021, 39(3): 615-637. DOI: 10.1109/JSAC.2020.3019724
[2] MIUCCIO L, PANNO D, RIOLO S. An energy-efficient DL-aided massive multiple access scheme for IoT scenarios in beyond 5G networks [J]. IEEE Internet of Things journal, 2023, 10(9): 7936 – 7959. DOI: 10.1109/JIOT.2022.3231128
[3] VISWANATHAN H, MOGENSEN P E. Communications in the 6G era [J]. IEEE access, 2020, 8: 57063 – 57074. DOI: 10.1109/ACCESS.2020.2981745
[4] WEI F, CHEN W, WU Y P, et al. Toward 5G wireless interface technology: enabling nonorthogonal multiple access in the sparse code domain [J]. IEEE vehicular technology magazine, 2018, 13(4): 18 – 27. DOI: 10.1109/MVT.2018.2867280
[5] LIU L, YU W. Massive connectivity with massive MIMO: part I: device activity detection and channel estimation [J]. IEEE transactions on signal processing, 2018, 66(11): 2933 – 2946. DOI: 10.1109/TSP.2018.2818082
[6] DING J, QU D M, FENG M J, et al. Dynamic preamble-resource partitioning for critical MTC in massive MIMO systems [J]. IEEE Internet of Things journal, 2021, 8(20): 15361 – 15371. DOI: 10.1109/JIOT.2021.3064061
[7] RUSEK F, PERSSON D, LAU B K, et al. Scaling up MIMO: opportunities and challenges with very large arrays [J]. IEEE signal processing magazine, 2013, 30(1): 40 – 60. DOI: 10.1109/MSP.2011.2178495
[8] SHEN J C, ZHANG J, CHEN K C, et al. High-dimensional CSI acquisition in massive MIMO: Sparsity-inspired approaches [J]. IEEE systems journal, 2017, 11(1): 32 – 40. DOI: 10.1109/JSYST.2015.2448661
[9] LIU L, LARSSON E G, YU W, et al. Sparse signal processing for grant-free massive connectivity: a future paradigm for random access protocols in the Internet of Things [J]. IEEE signal processing magazine, 2018, 35(5): 88 – 99. DOI: 10.1109/MSP.2018.2844952
[10] DAI J S, LIU A, LAU V K N. FDD massive MIMO channel estimation with arbitrary 2D-array geometry [J]. IEEE transactions on signal processing, 66(10): 2584 – 2599. DOI: 10.1109/tsp.2018.2807390
[11] RAJORIYA A, RUKHSANA S, BUDHIRAJA R. Centralized and decentralized active user detection and channel estimation in mMTC [J]. IEEE transactions on communications, 70(3): 1759 – 1776. DOI: 10.1109/tcomm.2022.3141401
[12] JALAL A, ARVINTE M, DARAS G, et al. Robust compressed sensing MRI with deep generative priors [C]//Proc. 35th International Conference on Neural Information Processing Systems. ACM, 2021: 14938 – 14954. DOI: 10.5555/3540261.3541406
[13] SONG Y, SHEN L Y, XING L, et al. Solving inverse problems in medical imaging with score-based generative models [EB/OL]. (2021-11-15) [2024-08-30]. https://arxiv.org/abs/2111.08005v2
[14] MENG X M, KABASHIMA Y. Diffusion model based posterior sampling for noisy linear inverse problems [EB/OL]. (2022-11-20) [2024-12-11]. https://arxiv.org/abs/2211.12343v4
[15] CANDÈS E J, ROMBERG J K, TAO T. Stable signal recovery from incomplete and inaccurate measurements [J]. Communications on pure and applied mathematics, 2006, 59(8): 1207 – 1223. DOI: 10.1002/cpa.20124
[16] WU Y, ROSCA M, LILLICRAP T. Deep compressed sensing [C]//36th International Conference on Machine Learning. PMLR, 2019: 6850 – 6860
[17] AALI A, ARVINTE M, KUMAR S, et al. Solving inverse problems with score-based generative priors learned from noisy data [C]//Proc. 57th Asilomar Conference on Signals, Systems, and Computers. IEEE, 2023: 837 – 843. DOI: 10.1109/IEEECONF59524.2023.10477042
[18] ALKHATEEB A, LEUS G, HEATH R W. Compressed sensing based multi-user millimeter wave systems: how many measurements are needed? [C]//Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015: 2909 – 2913.. DOI: 10.1109/icassp.2015.7178503
[19] REZENDE D J, MOHAMED S, REZENDE D J, et al. Variational inference with normalizing flows [C]//Proc. 32nd International Conference on International Conference on Machine Learning - Volume 37. ACM, 2015: 1530 – 1538. DOI: 10.5555/3045118.3045281
[20] SONG Y, ERMON S. Generative modeling by estimating gradients of the data distribution [J]. 2019: 11918 – 11930. DOI: 10.5555/3454287.3455354
[21] SONG Y, ERMON S, SONG Y, et al. Improved techniques for training score-based generative models [C]//Proc. 34th International Conference on Neural Information Processing Systems. ACM, 2020: 12438 – 12448. DOI: 10.5555/3495724.3496767
[22] MENG X M, KABASHIMA Y. Quantized compressed sensing with score-based generative models [EB/OL]. (2022-11-02) [2024-12-11]. https://arxiv.org/abs/2211.13006v4
[23] BORA A, JALAL A, PRICE E, et al. Compressed sensing using generative models [C]//Proc. 34th International Conference on Machine Learning. ACM, 2017: 537 – 546. DOI: 10.5555/3305381.3305437
[24] HASAN M, HOSSAIN E, NIYATO D. Random access for machine-to-machine communication in LTE-advanced networks: issues and approaches [J]. IEEE communications magazine, 2013, 51(6): 86 – 93. DOI: 10.1109/MCOM.2013.6525600
[25] PRATAS N K, THOMSEN H, STEFANOVIĆ Č, et al. Code-expanded random access for machine-type communications [C]//Proc. IEEE Globecom Workshops. IEEE, 2012: 1681 – 1686. DOI: 10.1109/GLOCOMW.2012.6477838
[26] CHEN Z L, YU W. Massive device activity detection by approximate message passing [C]//Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017: 3514 – 3518. DOI: 10.1109/ICASSP.2017.7952810
[27] DONOHO D L, MALEKI A, MONTANARI A. Message-passing algorithms for compressed sensing [J]. Proceedings of the national academy of sciences of the United States of America, 2009, 106(45): 18914 – 18919. DOI: 10.1073/pnas.0909892106
[28] BAYATI M, MONTANARI A. The dynamics of message passing on dense graphs, with applications to compressed sensing [J]. IEEE transac-

tions on information theory, 2011, 57(2): 764 – 785. DOI: 10.1109/TIT.2010.2094817
[29] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models [C]// 34th International Conference on Neural Information Processing Systems. NIPS, 2020: 6840 – 6851
[30] BATZOLIS G, STANCZUK J, SCHÖNLIEB C B, et al. Conditional image generation with score-based diffusion models [EB/OL]. (2021-11-26) [2024-12-11]. https://arxiv.org/abs/2111.13606v1
[31] SOLTANI M, POURAHMADI V, MIRZAEI A, et al. Deep learning-based channel estimation [J]. IEEE communications letters, 2019, 23(4): 652 – 655. DOI: 10.1109/LCOMM.2019.2898944
[32] HE H T, WEN C K, JIN S, et al. Deep learning-based channel estimation for beamspace mmWave massive MIMO systems [J]. IEEE wireless communications letters, 2018, 7(5): 852 – 855. DOI: 10.1109/LWC.2018.2832128
[33] METZLER C A, MOUSAVI A, BARANIUK R G, et al. Learned D-amp [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. ACM, 2017: 1770 – 1781. DOI: 10.5555/3294771.3294940
[34] ARVINTE M, TAMIR J I. MIMO channel estimation using score-based generative models [J]. IEEE transactions on wireless communications, 2023, 22(6): 3698 – 3713. DOI: 10.1109/TWC.2022.3220784
[35] LARSSON E G, EDFORS O, TUFVESSON F, et al. Massive MIMO for next generation wireless systems [J]. IEEE communications magazine, 52 (2): 186 – 195. DOI: 10.1109/mcom.2014.6736761
[36] BORGES D, MONTEZUMA P, DINIS R, et al. Massive MIMO techniques for 5G and beyond: opportunities and challenges [J]. Electronics, 2021, 10(14): 1667. DOI: 10.3390/electronics10141667
[37] TARANTOLA A. Inverse problem theory and methods for model parameter estimation [M]. Philadelphia, USA: Society for Industrial and Applied Mathematics, 2005. DOI: 10.1137/1.9780898717921
[38] BOTTARELLI M, EPIPHANIOU G, BEN ISMAIL D K, et al. Physical characteristics of wireless communication channels for secret key establishment: a survey of the research [J]. Computers & security, 2018, 78: 454 – 476. DOI: 10.1016/j.cose.2018.08.001
[39] VINCENT P. A connection between score matching and denoising autoencoders [J]. Neural computation, 2011, 23(7): 1661 – 1674. DOI: 10.1162/neco_a_00142
[40] SONG Y, DURKAN C, MURRAY I, et al. Maximum likelihood training of score-based diffusion models [EB/OL]. (2021-01-22) [2024-08-30]. https://arxiv.org/abs/2101.09258v4
[41] ROBERTS G O, TWEEDIE R L. Exponential convergence of Langevin distributions and their discrete approximations [J]. Bernoulli, 1996, 2(4): 341. DOI: 10.2307/3318418
[42] DANUFANE F H, DI RENZO M, DE ROSNY J, et al. On the path-loss of reconfigurable intelligent surfaces: an approach based on Green's theorem applied to vector fields [J]. IEEE transactions on communications, 2021, 69(8): 5573 – 5592. DOI: 10.1109/TCOMM.2021.3081452
[43] DATTA T, KUMAR N A, CHOCKALINGAM A, et al. A novel Monte-Carlo-sampling-based receiver for large-scale uplink multiuser MIMO systems [J]. IEEE transactions on vehicular technology, 2013, 62(7): 3019 – 3038. DOI: 10.1109/TVT.2013.2260572
[44] AIT AOUDIA F, HOYDIS J. End-to-end learning for OFDM: from neural receivers to pilotless communication [J]. IEEE transactions on wireless communications, 2022, 21(2): 1049 – 1063. DOI: 10.1109/TWC.2021.3101364

Biographies

TANG Chenyue (chenyue.23@intl.zju.edu.cn) received her BE degree from Central South University, China in 2023. She is currently working toward her master's degree in electrical engineering and information technology at Zhejiang University, China. Her current research interests include statistical signal processing, inverse problems, and machine learning.

LI Zeshen received his BE degree from Jilin University, China in 2022. He is working toward his PhD degree in information and communication engineering at Zhejiang University, China. His current research interests include federated learning, edge computing, and distributed machine learning.

CHEN Zihan received his BE degree in communication engineering from the Yingcai Honors College, University of Electronic Science and Technology of China (UESTC) in 2018 and PhD degree from the Singapore University of Technology and Design (SUTD)-National University of Singapore (NUS) Joint PhD Program in 2022. Currently, he is a postdoctoral research fellow with SUTD. His research mainly focuses on network intelligence, machine learning, and edge computing.

Howard H. YANG received his BE degree in communication engineering from Harbin Institute of Technology, China in 2012, and MSc degree in electronic engineering from the Hong Kong University of Science and Technology, China in 2013, and PhD degree in electrical engineering from the Singapore University of Technology and Design, Singapore in 2017. He is currently an assistant professor with Zhejiang University/University of Illinois at Urbana-Champaign Institute, Zhejiang University, China. His research interests currently focus on the modeling of modern wireless networks, high dimensional statistics, graph signal processing, and machine learning.

Efficient PSS Detection Algorithm Aided by CNN

LI Lanlan
(Shanghai Technical Institute of Electronics & Information, Shanghai 201141, China)

DOI: 10.12142/ZTECOM.202501008
https://kns.cnki.net/kcms/detail/34.1294.TN.20250220.1705.002.html, published online February 21, 2025
Manuscript received: 2025-01-10

Abstract: In a 5G mobile communication system, cell search is the initial step in establishing downlink synchronization between user equipment (UE) and base stations (BS). Primary synchronization signal (PSS) detection is a crucial part of this process, and enhancing PSS detection speed can reduce communication latency and improve overall quality. This paper proposes a fast PSS detection algorithm based on the correlation characteristics of PSS time-domain superposition signals. Conducting PSS signal correlation within a smaller range can reduce computational complexity and accelerates communication speed. Additionally, frequency offset can impact the accuracy of calculations during the PSS detection process. To address this issue, we propose applying convolutional neural networks (CNN) for frequency offset estimation of synchronization signals. By compensating for the frequency of related signals, the accuracy of PSS detection is improved. Finally, the analysis and simulation results demonstrate the effectiveness of the proposed approach.

Keywords: 5G; CNN; cell search; PSS detection

Citation (Format 1): LI L L. Efficient PSS detection algorithm aided by CNN [J]. ZTE Communications, 2025, 23(1): 63 - 70. DOI: 10.12142/ZTECOM.202501008
Citation (Format 2): L. L. Li, "Efficient PSS detection algorithm aided by CNN," *ZTE Communications*, vol. 23, no. 1, pp. 63 - 70, Mar. 2025. doi: 10.12142/ZTECOM.202501008.

1 Introduction

The communication between user equipment (UE) and the base station (BS) is established via wireless signals, where cell search serves as the initial step for terminal devices to access the 5G network. After power on, users need to perform a cell search to quickly identify their current cell, obtain the cell ID, and achieve time-frequency synchronization. The detection of the primary synchronization signal (PSS) in 5G is an important process in wireless communication, involving the identification and decoding of the PSS from received signals. PSS is one of the signals that help the UE synchronize with base stations, enabling devices to determine the start of wireless frames and decode additional signals. It also plays a key role in identifying 5G cells; when combined with the secondary synchronization signal (SSS), it can uniquely identify a cell. Through effective algorithms and techniques, efficient PSS detection can be achieved under various channel conditions, ensuring reliable synchronization and network access capabilities for the UE.

Several research findings regarding PSS detection algorithms have been proposed. Starting from the basic principles of the 5G initial access process, JEON et al.[1] proposed the cell search process and the PSS structure of the 5G communication system. CHAKRAPANI[2] proposed the composition of the synchronization signal block (SSB) carrying PSSes. BALASUBRAMANYA et al.[3] proposed a design scheme for 4G PSS in the evolution of 5G technology. A new method for rapid detection of PSS by UE was introduced in Ref. [4], which improved fast synchronization between terminals and networks. YOU[5] proposed a sequential integer carrier frequency offset (ICFO) and edge master synchronization signal (S-PSS) detection scheme to reduce complexity in the 5G new wireless vehicular Internet of Things system. There are also various solutions to the frequency offset problem in PSS detection[6]. In Ref. [7], the authors described a program for synchronizing 5G networks and proposed two methods to estimate frequency offset (FFO). The first method utilizes the carried information, and the second method involves partial cross-correlation of PSS, which is applied to each orthogonal frequency division multiplexing (OFDM) symbol in the SSB, with the phase of the auto-correlation peak used to estimate the value of FFO. However, synchronization errors can reduce the performance of maximum likelihood (ML) methods[8]. Some researchers have adopted joint detection and estimation methods for initial downlink access, as described in Refs. [9] and [10]. The second technique for estimating FFO is based on replicating the correlation signal between the partial input PSS and the PSS over more than half of the symbol length duration[11]. It is also noted that synchronization errors can diminish the performance of the FFO estimation method. In the past two years, significant advancement has been made in 5G PSS detection and the application of convolutional neural networks

(CNN) in physical layer algorithms. ASSAF et al.[12] evaluated 5G New Radio (NR) frequency synchronization in the downlink initial access, and proposed and investigated a reduced-complexity FFO estimation method. In Ref. [13], a novel approach to enhancing the detection of PSS sequences in 5G NR systems was proposed. ZHANG et al.[14] proposed a scheme to estimate the energy per resource element (EPRE) ratio of PSS to SSS/demodulation reference signal (DMRS) and demonstrated the proposed scheme can estimate the EPRE ratio accurately when the signal-to-noise ratio (SNR) is above −4 dB through simulation results. COUTINHO et al.[15] proposed a CNN-based algorithm for channel estimation in the presence of phase noise and carrier frequency offset (CFO) in 5G and beyond systems. ZHENG et al.[16] proposed a decomposed CNN for the sub-Nyquist tensor-based 2D direction of arrival (DoA) estimation.

The main motivation and novelties of this paper are summarized as follows.

• This paper proposes a fast PSS detection algorithm assisted by a CNN neural network, which can quickly complete the PSS detection process after the 5G terminal device is turned on, thereby reducing communication latency.

• In the fast PSS detection algorithm, the sum sequence, obtained by superimposing three frequency domain PSS sequences, is cross-correlated with the received signal in the time domain. A shorter time-domain sequence is determined based on the correlation peak and then transformed into the frequency domain to cross-correlate with the received signal. The cell ID required for PSS detection is determined from the correlation peak.

• Local received signals typically have a frequency offset. Using CNN-assisted frequency offset correction algorithms can yield corrected received signals, thereby enhancing the accuracy of PSS detection results.

2 Background Description

2.1 5G Cell Search Procedure

The 5G NR cell search process is a key step for UE to find and access suitable serving cells in the network when it is turned on or needs to reconnect. The specific steps of the 5G NR cell search are as follows:

Step 1: The NR terminal adjusts the radio frequency (RF) receiver to the designated receiving frequency to capture the signal;

Step 2: The PSS synchronization detection is performed to obtain time slot timing information and retrieve the sector number $N_{ID}^{(2)}$ within the cell group;

Step 3: Frequency offset compensation is applied;

Step 4: Based on the relationship between PSS and SSS in the synchronization signal and the physical broadcast channel (PBCH) block, the NR terminal performs frequency domain correlation detection on the SSS to obtain the cell group number $N_{ID}^{(1)}$;

Step 5: The NR terminal obtains the cell ID using the previously obtained cell group ID $N_{\mathrm{ID}}^{(2)}$ and cell group ID $N_{\mathrm{ID}}^{(1)}$. Then, retrieve the corresponding DMRS information from the PBCH based on the cell ID to obtain the SSB index, which corresponds to the beam ID[17];

Step 6: The PBCH symbol is decoded to obtain the master information block (MIB) information;

Step 7: The cell search process is completed, enabling the UE to perform a random signal access process for uplink synchronization.

2.2 PSS Detection

From the cell search process described above, it is evident that the PSS synchronization detection process is the initial step for mobile terminals to access the network. This step enables terminal devices to perform tasks such as sector identification $N_{\mathrm{ID}}^{(2)}$ recognition, frequency synchronization, neighbor cell search, and fast locking. Specifically, after several steps, such as coarse time synchronization, frequency offset estimation, fine synchronization, SSS detection, and beam ID detection, users can receive and interpret the physical broadcast information of the cell, obtain MIB and system information block (SIB), and complete cell access through random access and other processes based on the system messages received. In these steps, coarse time synchronization involves positioning the timing synchronization within the cyclic prefix range, which is accomplished using PSS signals. 5G PSS has strong autocorrelation and cross-correlation properties, which are leveraged for coarse time synchronization. Since there are only three sets of PSS sequences and the generation of SSS signals is linked to both cell group identification and sector identification, performing PSS detection first reduces synchronization complexity and facilitates the retrieval of necessary physical cell information. By utilizing the correlation characteristics of the PSS to demodulate the PSS in the received signal, the starting position of OFDM symbols and the sector ID, $N_{\mathrm{ID}}^{(2)}$, carried by PSS can be determined. Based on the fixed time-frequency position of SSB, once the time-frequency position of PSS is established, the time-frequency position of SSS can be determined. The frequency domain position of the SSS matches that of the PSS, while in the time domain, the SSS is shifted by two OFDM symbols from the position of the PSS. Using the generation rules or cross-correlation characteristics of SSS, the SSS sequence can be demodulated to determine the cell group ID, $N_{\mathrm{ID}}^{(1)}$, carried by SSS. The cell identification number can be calculated from the relationship between the cell group ID and the sector ID, completing the downlink synchronization process and allowing the terminal to access the base station's network. From the above process, it is clear that quickly determining the frequency domain position of PSS can improve the speed of cell search, enabling terminals to access the network more rapidly.

The traditional PSS detection algorithm generates a local

PSS time-domain sequence and performs cross-correlation calculations with the received signal. The PSS sequence has good correlation characteristics, and sliding cross-correlation can fully leverage these properties.

First, three sets of local PSS time-domain signals are generated, followed by point-by-point sliding cross-correlation with the received signal. Significant peaks occur only when the local PSS sequence matches the PSS sequence in the received signal. The maximum correlation value is identified, and the position of this maximum value serves as the synchronization point for the PSS. Simultaneously, the PSS sequence that detects the peak corresponds to the sector ID number it carries.

The sliding cross-correlation detection process is shown in Fig. 1.

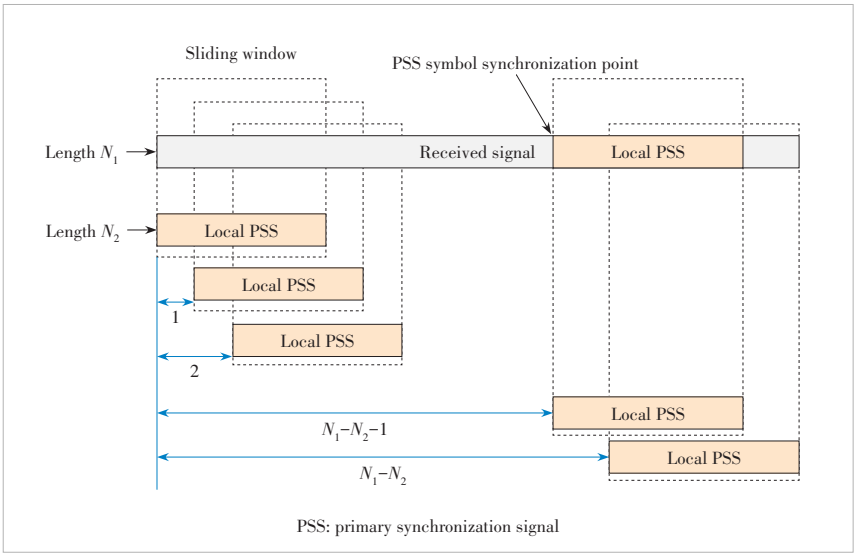


Figure 1. Traditional PSS detection algorithm

The frequency band occupied by 5G communications is relatively broad, encompassing a total of 29 frequency bands. They are primarily divided into two spectrum ranges: 26 frequency bands below 6 GHz (collectively referred to as sub-6 GHz) and 3 millimeter wave frequency bands. Currently, sub-6 GHz is primarily used in China, and it includes 7 frequency bands: n1, n3, n28, n41, n77, n78, and n79. 5G supports a maximum bandwidth configuration of 400 MHz. In the standalone (SA) mode, the SSB frequency domain location where the PSS is located must be determined by the global synchronization channel number (GSCN). Due to the extensive bandwidth of 5G NR, the concepts of GSCN and the Global Synchronization Grid have been introduced. The SSB frequency domain is positioned at integer intervals of the Global Synchronization Grid and terminals search for synchronization signals at these intervals. For frequencies below 3 GHz, the frequency scanning interval is 1.2 MHz; for frequencies between 3 GHz and 24.25 GHz, the interval is 1.44 MHz; for frequencies between 24.25 GHz and 100 GHz, the scanning interval is 17.28 MHz. The frequency range, SSB position, and GSCN determination are outlined in Table 1.

Table 1. Global synchronization grid

Frequency Range/MHz	SSB	GSCN	GSCN Range
0 – 3 000	N*1 200 kHz+M*50 kHz, N=1:2 499, $M\in(1,3,5)$	$3N+(M-3)/2$	2 – 7 498
3 000 – 24 250	3 000 MHz+N*1.44 MHz, N=0:14 756	7 499+N	7 498 – 22 255

GSCN: global synchronization channel number
SSB: synchronization signal block

In the non-standalone (NSA) mode, the SSB frequency domain position is also uncertain, and the terminal is notified of the SSB frequency point position through high-level signaling. This introduces uncertainty in the SSB position across the entire bandwidth. The PSS sequence is a part of the SSB, as shown in Fig. 2, and the frequency-domain position of the PSS sequence is similarly uncertain across the entire bandwidth.

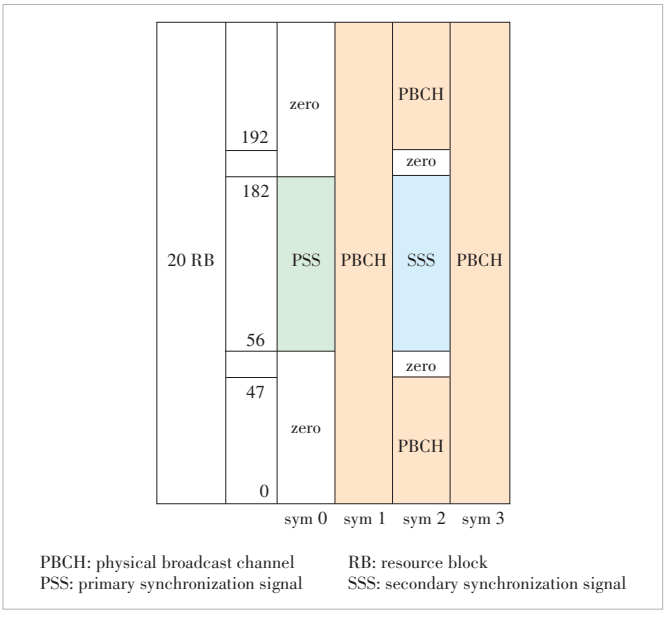


Figure 2. Structure of synchronization signal block

PSS sequences at different frequency domain positions may generate distinct time-domain sequences through the inverse fast Fourier transform (IFFT), leading to a rapid increase in computational complexity, which is unsuitable for 5G NR systems. Additionally, the large volume of received data further exacerbates computational complexity. This combination results in higher computational complexity, causing significant computation delays, longer communication delays, and reduced network communication quality. To address these challenges, this paper proposes a CNN-assisted PSS detection method to quickly determine the frequency domain position of

the PSS, thereby shortening synchronization time and accelerating the cell search process. The existing PSS signal synchronization detection algorithm performs correlation operations on PSS sequences at various frequency points within the working frequency band in the time domain. Due to the lengthy PSS sequence (and consequently, the received signal), correlating the three PSS sequences with the received signal leads to high algorithm complexity and considerable computational demands, resulting in prolonged communication delays. Moreover, unlike 4G technology, the SSB in 5G NR is no longer fixed in the middle of the frequency band. The flexible placement of SSB time-frequency positions increases the initial blind detection computation of PSS, impacting the speed at which users can decode base station broadcast information and ultimately diminishing network communication quality.

2.3 System Model

During propagation, the transmitted signal is first corrupted by multi-path fading and additive white Gaussian noise (AWGN). CFO is introduced owing to the oscillator mismatch between BS and UE. The received signal is then modeled as[19]:

$$r(n) = s(n)e^{j\frac{2\pi\varepsilon n}{N}} + \omega(n) \quad (1),$$

where $s(n)$ is the transmitted signal, $\omega(n)$ is the zero mean AWGN with unity variance, and ε denotes the relative CFO normalized by the sub-carrier frequency spacing.

3 CNN-Assisted Fast PSS Detection Algorithm

Given the high complexity of traditional PSS detection algorithms and their limited resistance to frequency offset and noise[20], there is a pressing need for a new algorithm that offers fast processing speed, anti-frequency offset capabilities, and effective correlation utilization. To address this, this paper proposes an algorithm based on the CNN method to process the received signal sequence in the presence of frequency offset. It further leverages the cross-correlation features of frequency-domain superimposed signals to optimize PSS detection. This approach not only enhances the resistance to frequency offset but also significantly improves the PSS detection speed, thereby reducing communication latency. The processing flow of the proposed PSS detection optimization algorithm consists of the following steps.

1) Step 1: Generate a polynomial based on the PSS sequence. The specific implementation method is as follows. There are 1 008 physical layer cells in NR, and the formula for calculating NR cell IDs is:

$$N_{\mathrm{ID}}^{\mathrm{cell}} = 3N_{\mathrm{ID}}^{(1)} + N_{\mathrm{ID}}^{(2)} \quad (2),$$

where $N_{\mathrm{ID}}^{(1)} \in \{0,1,\cdots,335\}$, carried by SSS, and $N_{\mathrm{ID}}^{(2)} \in \{0,1,2\}$, carried by the PSS. The primary synchronization signal is defined in 3GPP protocol TS38.211 and utilizes three m-sequences of length 127 to represent the three values of $N_{\mathrm{ID}}^{(2)}$.

To construct the PSS sequence, zeros are inserted at both ends of the $d_{\mathrm{PSS},i}(k)$ sequence (where i = 0,1,2 and k= 56, 57,⋯,182) for a local sequence length of 127. This process extends the sequence to a total length of 256, resulting in $\mathrm{PSS}_i(k)$, which is expressed as:

$$\mathrm{PSS}_i(k) = \begin{cases} 0, & k = 0,1,2,\cdots,55,183,\cdots,255 \\ d_{\mathrm{PSS},i}(k), & k = 56,57,\cdots,182 \end{cases} \quad (3),$$

where i=0, 1, and 2. The generation formula maps $\mathrm{PSS}_i(k), i \in \{0,1,2\}$ to the corresponding $N_{\mathrm{ID}}^{(2)}$.

2) Step 2: Overlay three frequency-domain PSS sequences. In the second step, the three frequency-domain PSS sequences are overlaid to create a sum sequence $\mathrm{PSS}_{\mathrm{sum}}$. An IFFT is then applied to convert the frequency-domain sequence into a time-domain sequence $\mathrm{pss\_}t_{\mathrm{sum}}(k), k$ = 0,1,2,⋯,255. The specific implementation method is as follows. Denote the three frequency-domain PSS sequences as $\mathrm{PSS}_i(k)$, where i = 0,1,2. We compute the element-wise sum of the three sequences to obtain the sum sequence $\mathrm{PSS}_{\mathrm{sum}}(k)$ and represent it as:

$$\mathrm{PSS}_{\mathrm{sum}}(k) = \sum_{i=0}^{2} \mathrm{PSS}_i(k), k = 0,1,2,\cdots,255 \quad (4).$$

The sequence shown above is transformed from a frequency domain sequence to a time domain sequence $\mathrm{pss\_}t_{\mathrm{sum}}(k)$ through the IFFT process, which can be expressed as:

$$\mathrm{pss\_}t_{\mathrm{sum}}(k) = \mathrm{IFFT}\left(\mathrm{PSS}_{\mathrm{sum}}(k)\right), k = 0,1,2,\cdots,255 \quad (5).$$

3) Step 3: Estimate signal reception and frequency offset using CNN. In the third step, the terminal receives the time-domain signal $\tilde{r}(k)$ transmitted by the base station. A CNN model is then employed to correct the received signal and estimate the carrier frequency offset, yielding $r(k)$. The CNN-based carrier frequency offset estimation consists of two stages: offline training and online estimation. Firstly, the offline training process involves generating a network training dataset through MATLAB simulation based on the statistical characteristics of the signal used for frequency offset estimation. The dataset is processed from complex to real numbers and then used for offline training of the model. Finally, the trained network model parameters are saved. When estimated online, the received OFDM system signal $\tilde{r}(k)$ is converted into real numbers and transmitted to the trained CNN model. The estimation result $r(k)$ can be directly output based on the trained network parameters.

4) Step 4: Determine the peak value and time offset using correlation operation. A correlation operation is performed between the sequence and the time-domain received signal $r(k)$ to determine the peak value and corresponding time offset value k_0. The specific implementation process is as follows.

• Cross-correlation operation

We cross-correlate the time-domain sequence $\text{pss\_}t_{\text{sum}}(k), k = 0,1,2,\cdots,255$ with the local received signal $r(k)$, where k=0, 1, 2,⋯, 255. The cross-correlation function $C(k)$ is defined as:

$$C(k) = \left| \sum_{n=0}^{N-1} \text{pss\_}t^{*}_{\text{sum}}(n) r(k+n) \right|^2 \tag{6}$$

Here, $\text{pss\_}t^{*}_{\text{sum}}(n)$ is the complex conjugate of $\text{pss\_}t_{\text{sum}}(n)$, and N is the length of the sequence.

• Synchronization position determination

The position k_0 corresponding to the maximum value of the correlation peak is calculated as :

$$k_0 = \arg\max_{k} \{C(k), k = 0,1,2,\cdots\} \tag{7}$$

• Visualization of cross-correlation results

Fig. 3 illustrates the cross-correlation results among the time-domain received signals, the three local time-domain PSS sequences, and their superimposed and constructed sequences. The time-domain signals are obtained by applying an IFFT to the frequency-domain representations of the PSS sequences and their superposition. These time-domain signals are then cross-correlated with the received signal to calculate their correlation peak values.

• Analysis of correlation peak results

From Fig. 3, it is evident that the correlation peak values of the superimposed sequence $\text{PSS}_{\text{sum}}(k)$ in the time domain align with the trend of the correlation peak values of the individual PSS sequences, e.g., $\text{pss\_}t_3(k)$. While the peak magnitude of $\text{PSS}_{\text{sum}}(k)$ is slightly lower than that of a specific PSS sequence, and the difference is negligible. This demonstrates the feasibility of using the superimposed PSS to determine correlation peak values and derive the corresponding time offset k_0.

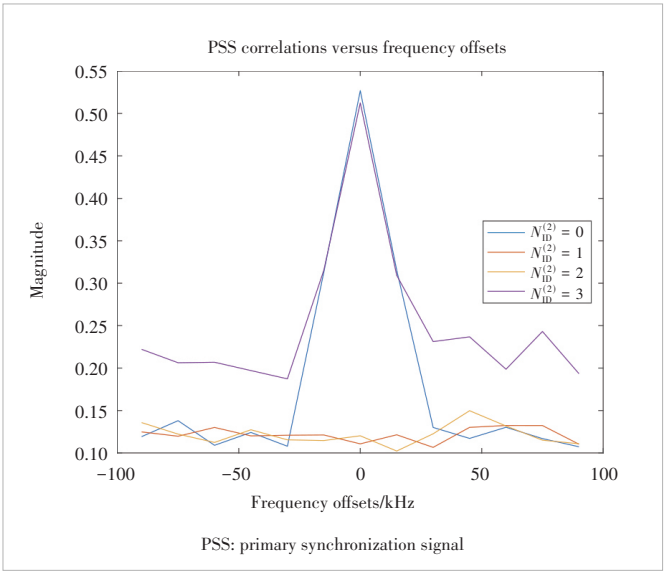


Figure 3. Correlation peaks of superimposed signals

• Example of cell ID correlation

Fig. 3 shows the three time-domain sequences $\text{pss\_}t_i(k), i = 1,2,3$, where i=1, 2, and 3 correspond to cell IDs 1, 2, and 3, respectively. The correlation results confirm that the superimposed sequence can reliably achieve time-domain synchronization for these cell IDs.

5) Step 5: Extract and transform the time-domain signal to frequency domain. In this step, a portion of the time-domain received signal is extracted from the corresponding time offset position k_0 to obtain a shorter time-domain signal sequence. The signal is then transformed into the frequency domain using FFT to obtain the frequency-domain signal segment $R0(k)$. The specific implementation method is as follows.

• Signal extraction

Starting from the corresponding time offset position k_0, we intercept a segment of the time-domain signal $r(k)$. The extracted signal segment is denoted as $r0(k)$, and its length corresponds to the OFDM symbol length L that depends on the number of sampling points, represented as intercept(k).

• Frequency-domain transformation and output

The extracted frequency domain representation of the received signal is obtained as $R0(k), k = 1,2,\cdots,L$. The signal $r0(k)$ is transformed by FFT into the frequency domain signal $R0$, denoted as $R0(k)$, where $R0(k) = \text{FFT}(r0(k)), k = 1,2,\cdots,L$.

6) Step 6: Perform correlation to determine the PSS sequence ID. Here, the received signal is correlated with the three possible PSS sequences $\text{PSS}_i, i = 1,2,3$, to determine the ID of the PSS sequence. The specific implementation method is as follows. The frequency domain signal $R0(k)$ is then correlated with three local frequency domain sequences $\text{PSS}_i(k), i = 0,1,2$. The maximum peak of the correlation value for each possibility of i is taken, and these three correlation values are compared to obtain the maximum value. Based on the corresponding frequency domain signal $\text{PSS}_i(k), i = 0,1,2$, the corresponding small cell group number N_{ID}^{2} can be obtained, and the corresponding PSS sequence ID can be further determined. The mathematical expression for the above process is:

$$\text{corr}_i = \sum_{n} R0(n+k) + \text{PSS}_i(n) \tag{8}$$

$$\text{PSS}_{id} = \max_{i} (\text{abs}(\text{corr}_i)) \tag{9}$$

4 Simulation and Analysis

To evaluate the performance of the proposed PSS search algorithm, a 5G cell search link was constructed using MATLAB 2021a. The channel environment was modeled using the tapped delay line-A (TDL-A) model channel. The

simulation parameters for cell search are shown in Table 2. This section simulates the main synchronization process of the 5G NR system using MATLAB.

The simulation steps are as follows. First, according to 3GPP TS38.211[18], a downlink signal containing SSB is generated for a cell with a cell identifier of 2 ($N_{\mathrm{ID}}^2 = 2$), using parameters in Table 2. Next, the generated signal is passed through a channel model to simulate the received signal. The 5G NR channel model used in the simulation is a TDL. Finally, different PSS detection algorithms are applied using the received 5G signals for performance evaluation.

Fig. 4 shows the peak values obtained using the proposed algorithm under the aforementioned simulation conditions. The three subgraphs are calculated using the three local sets $\{N_{\mathrm{ID}}^2, \mathrm{ID} \in (0,1,2)\}$ of PSS. The proposed algorithm successfully identifies the correct N_{ID}^2 and PSS synchronization points.

Fig. 5 shows the comparison of PSS detection results between the improved algorithm and the existing algorithm with different frequency offset parameters. The accuracy of PSS detection by the improved algorithm is higher than that of the existing algorithm. Especially, when the frequency offset is large, the PSS detection accuracy of the improved algorithm is significantly improved compared with existing algorithms. The proposed superimposed cross-correlation method can mitigate the frequency offset accumulation of sliding cross-correlation. Combined with the CNN method for frequency offset correction of the received signal, it offers better detection performance and lower computational complexity than the traditional sliding cross-correlation method.

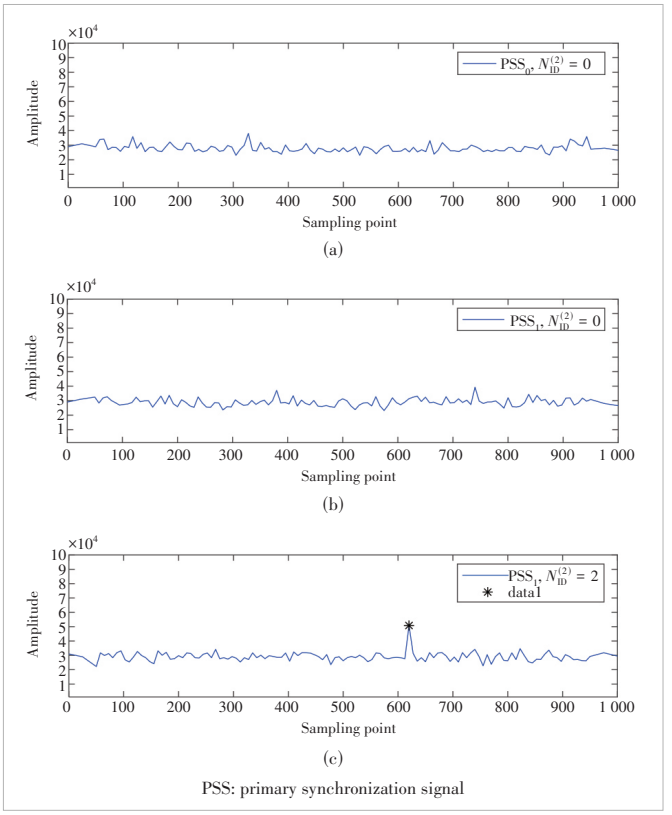
Fig. 6 shows when the SNR is low, the time consumption difference between the proposed algorithm and the baseline algorithm is not significant; on the contrary, when the SNR is high, using the proposed algorithm to perform PSS detection takes much less time than the baseline algorithm, indicating that the proposed algorithm is more suitable for scenarios with high SNRs.

Fig. 7 illustrates the accuracy of PSS synchronization under various frequency offsets. As the frequency offset increases,

Table 2. Simulation parameters for cell search

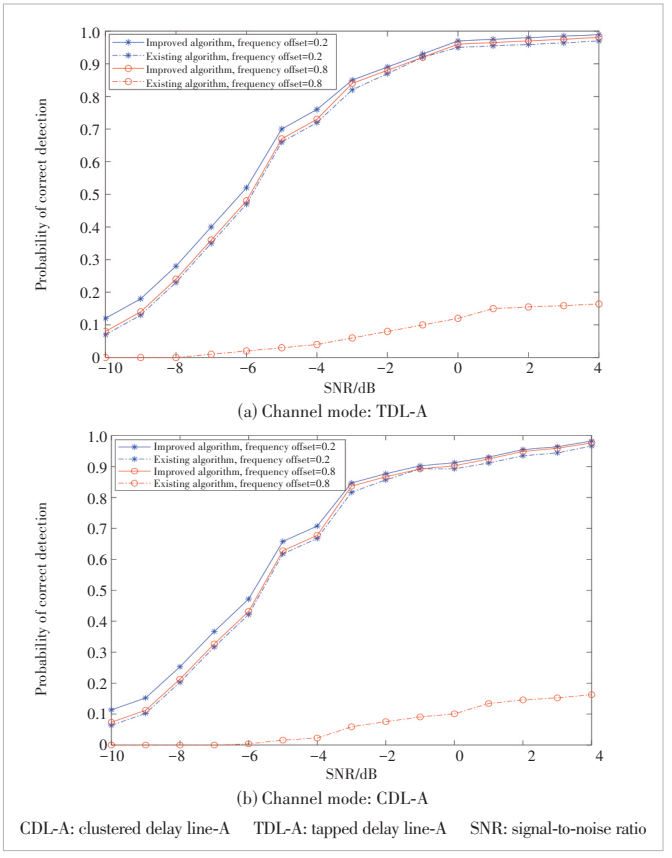
Simulation Parameter Types	Configuration Parameters
Channel bandwidth/MHz	100
Subcarrier spacing/kHz	15, 30
The number of FFT points	1 024, 4 096
Channel mode	TDL-A, CDL-A
Sampling frequency/MHz	122.88
Frequency offset/kHz	0.2, 0.8, 2.8
SSB block type	Case C
CP type	Standard

CDL-A: clustered delay line-A
CP: cyclic prefix
FFT: fast Fourier transform
SSB: synchronization signal block
TDL-A: tapped delay line-A



PSS: primary synchronization signal

Figure 4. Correlation peak plot calculated by proposed algorithm



CDL-A: clustered delay line-A TDL-A: tapped delay line-A SNR: signal-to-noise ratio

Figure 5. Probability of primary synchronization signal search algorithm

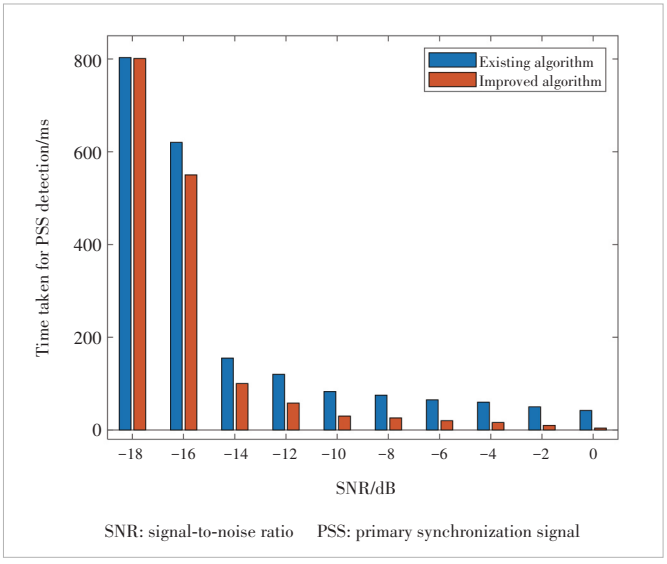


Figure 6. Time consumption difference between proposed algorithm and existing algorithm

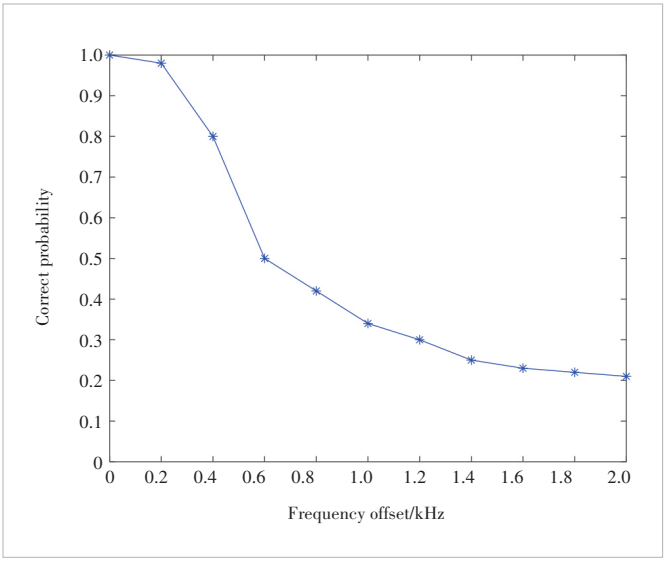


Figure 7. Accuracy of synchronization under different frequency offsets

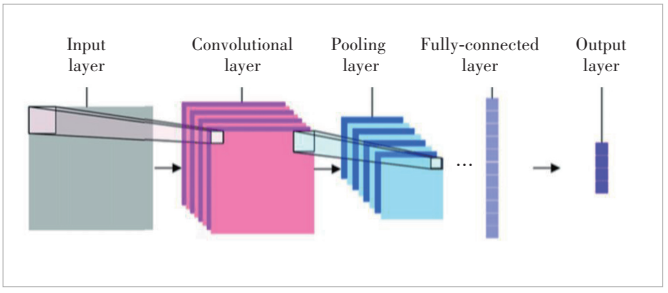


Figure 8. Basic structure of convolutional neural networks

the algorithm performs well and tends to stabilize.

Fig. 8 illustrates the CNN neural network architecture, which consists of several key layers designed to optimize PSS detection in 5G NR systems. The architecture begins with an input layer that processes the received signal data, followed by a series of convolutional layers that extract relevant features from the signal. Each convolutional layer is paired with activation functions, such as the rectified linear unit (ReLU), to introduce non-linearities. These layers are followed by pooling layers that reduce the dimensionality of the feature maps, which decreases computational complexity and improves generalization. The final layers include fully connected layers that aggregate the features and output a classification decision or prediction, such as the PSS sequence's position or the sector ID. This CNN architecture is tailored to enhance detection accuracy and robustness against frequency offsets and noise, making it suitable for high-performance PSS detection in dynamic 5G environments.

In the conventional method, the main complexity comes from the correlation operations, while in our proposed method, it comes from correlation operations and convolution layers in the CNN block. Unlike existing algorithms, our proposed algorithm exhibits higher complexity, primarily due to the operations of the CNN. Suppose the length of a data frame is L. After downsampling, the length of the received signal is K. The length of a downsampling time-domain PSS sequence is N. Using the traditional sliding correlation method, the sliding window length is $K-N+1$, representing the number of correlations required for a set of local PSS signals to complete synchronization detection. Each correlation operation involves N complex multiplications and $N-1$ complex additions. Therefore, sliding cross-correlation requires $3N(K-N+1)$ complex multiplications and $3(N-1)(K-N+1)$ complex additions. The order of magnitude of the calculation is $3O(NK)$. The proposed superimposed correlation method requires $N(K-N+1)+3N(L-N+1)$ complex multiplications and $N(K-N+1)+2N+3N(L-N+1)$ complex additions, where $L \ll K$. Given P is the number of transmitting antennas, M is the number of receiving antennas, and N_c is the number of subcarriers, with the CNN network comprising two convolutional layers of kernel size 3 (see Fig. 8), the additional complexity introduced by the algorithm is $O\left(2P \times M \times 2 \times 3^2\right)$. The proposed algorithm enhances detection and estimation performance, especially in the presence of a CFO. Considering the computational load of the CNN algorithm, the order of magnitude of the calculation is $O(NK)$. The total computational complexity is less than that of the traditional sliding correlation method.

Integrating AI modules and data processing units into 5G base stations enables the implementation of AI-related algorithms. This architecture can be guided by relevant patents[21].

5 Conclusions

This paper analyzes existing PSS synchronization detection algorithms and their characteristics in 5G NR systems, verifying the relationship between the autocorrelation peaks and frequency offset of three superimposed PSS signals compared

with a single PSS signal through experimental results. The accuracy of the CNN-assisted frequency offset estimation algorithm is examined, leading to the proposal of a new fast PSS synchronization detection algorithm that offers resistance to frequency offset and noise. In the cell search process, a method is introduced to determine a shorter synchronization signal sequence based on the frequency domain offset consistency between the autocorrelation peak of the superimposed PSS signals and the correlation peak of non-superimposed signals. This approach reduces the computational load of PSS synchronization detection and enhances the efficiency of the NR communication system's cell search. The simulation results demonstrate that the improved algorithm effectively enhances synchronization detection performance under large CFO conditions in the TDL-A or CDL-A channel. Future research will focus on developing PSS detection algorithms suitable for low SNR scenarios. The performance of the CNN model in highly dynamic or interference-heavy environments, along with the computational burden on terminals and the energy consumption of running CNN models on resource-constrained devices, will be studied in future research.

References

[1] JEON Y, PARK H, CHOI E. Synchronization and cell search procedure in 3GPP 5G NR systems [C]//The 21st International Conference on Advanced Communication Technology (ICACT). IEEE, 2019: 475 – 478
[2] CHAKRAPANI A. On the design details of SS/PBCH, signal generation and PRACH in 5G-NR [J]. IEEE access, 2020, 8: 136617 – 136637. DOI: 10.1109/ACCESS.2020.3010500
[3] BALASUBRAMANYA N M, LAMPE L, VOS G, et al. On timing reacquisition and enhanced primary synchronization signal (ePSS) design for energy efficient 3GPP LTE MTC [J]. IEEE transactions on mobile computing, 2016, 16(8): 2292 – 2305. DOI: 10.1109/TMC.2016.2618865
[4] VANKAYALA S K, AKHTAR J, ASHOK KRISHNAN K S, et al. Accelerated detection schemes for PSS in 5G-NR [C]//Proceedings of IEEE 3rd 5G World Forum (5GWF). IEEE, 2020. DOI: 10.1109/5gwf49715.2020.9221156
[5] YOU Y H. Reduced complexity frequency ambiguity resolution and synchronization signal detection for 5G NR-V2X communication systems [J]. IEEE systems journal, 2022, 16(1): 1325 – 1333. DOI: 10.1109/JSYST.2021.3058673
[6] LI G Q, XU Y H, LIN J Z, et al. Res-DNN based signal detection algorithm for end-to-end MIMO systems [J]. Chinese journal on Internet of Things, 2022, 6(1): 65 – 72. DOI: 10.11959/j.issn.2096-3750.2022.00256
[7] OMRI A, SHAQFEH M, ALI A, et al. Synchronization procedure in 5G NR systems [J]. IEEE access, 2019, 7: 41286 – 41295. DOI: 10.1109/ACCESS.2019.2907970
[8] VAN DE BEEK J J, SANDELL M, BORJESSON P O. ML estimation of time and frequency offset in OFDM systems [J]. IEEE transactions on signal processing, 1997, 45(7): 1800 – 1805. DOI: 10.1109/78.599949
[9] CHIU K L, SHEN P H, LIN B R, et al. Design of downlink synchronization for millimeter wave cellular system based on multipath division multiple access [J]. IEEE transactions on circuits and systems I: regular papers, 2020, 67(9): 3211 – 3223. DOI: 10.1109/TCSI.2020.2989467
[10] YIN J L, LEE M C, HSIAO W H, et al. A novel network resolved and mobile assisted cell search method for 5G cellular communication systems [J]. IEEE access, 2022, 10: 75331 – 75342. DOI: 10.1109/ACCESS.2022.3191357
[11] RICCARDO TUNINATO. Algorithms for New Radio synchronization layer functions [D]. Turin: Politecnico di Torino, 2020.
[12] ASSAF M, PONOMAREV O G. Efficient and low complexity frequency synchronization in NR-5G downlink [C]//The 25th International Conference on Digital Signal Processing and its Applications (DSPA). IEEE, 2023: 1 – 6. DOI: 10.1109/DSPA57594.2023.10113363
[13] KWAK B J, KIM J K, MYUNG J H, et al. A novel approach for enhancing 5G NR PSS detection with non-matched filter [C]//The 15th International Conference on Information and Communication Technology Convergence (ICTC). IEEE, 2024: 1642 – 1643. DOI: 10.1109/ICTC62082.2024.10827761
[14] ZHANG C L, WU Q, ZHAO X W, et al. A scheme for improving 5G cell search performance [C]//The 5th International Conference on Communications, Information System and Computer Engineering (CISCE). IEEE, 2023: 141 – 145. DOI: 10.1109/CISCE58541.2023.10142271
[15] COUTINHO F D L, SILVA H S, GEORGIEVA P, et al. A novel CNN-based channel estimation algorithm in the presence of phase noise and CFO [J]. IEEE wireless communications letters, 2024, 13(1): 193 – 197. DOI: 10.1109/LWC.2023.3325129
[16] ZHENG H, ZHOU C, VOROBOYV S A, et al. Decomposed CNN for sub-nyquist tensor-based 2-D DOA estimation [J]. IEEE signal processing letters, 2023, 30: 1708 – 1712
[17] ASSAF M, PONOMAREV O G. On DMRS tracking mode synchronization in fifth generation new radio [C]//Proceedings of 24th International Conference on Digital Signal Processing and its Applications (DSPA). IEEE, 2022: 1 – 6. DOI: 10.1109/DSPA53304.2022.9790773
[18] 3GPP. 3GPP technical specification group radio access network; physical channels and modulation: TS38.211 [S]. 2022
[19] WANG D, MEI Z Q, ZHANG H Q, et al. A novel PSS timing synchronization algorithm for cell search in 5G NR system [J]. IEEE access, 2021, 9: 5870-5880
[20] HE Y X, GU Y, BU S Q, et al. Primary synchronization signal design for New Radio technique in 5G communication system [EB/OL]. (2017-12-08)[2024-10-06]. http://dx.doi.org/10.4108/eai.13-7-2017.2270278
[21] LI L L, YI Y S, ZHANG C, et al. Interactive methods, devices, equipment, and storage media for wireless networks (in Chinese) [R]. Nanjing, China: Purple Mountain Laboratories, 2020

Biography

LI Lanlan (403197915@qq.com) received her MS degree in applied mathematics from Xinjiang University, China in 2001, and PhD degree in electrical engineering from Southeast University, China in 2005. She is currently engaged in teaching and research at the Shanghai Technical Institute of Electronics & Information, China. From 2019 to 2022, she worked as a researcher in intelligent communication at the Purple Mountain Laboratories, China. Her research interests include radio resource management, signal processing for digital communications, 6G communication technology, and AI-related applications in communication.

A Basis Function Generation Based Digital Predistortion Concurrent Neural Network Model for RF Power Amplifiers

SHAO Jianfeng[1], HONG Xi[1], WANG Wenjie[1], LIN Zeyu[2], LI Yunhua[2]
(1. Xi'an Jiaotong University, Xi'an 710049, China;
2. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202501009
https://kns.cnki.net/kcms/detail/34.1294.TN.20250224.1024.002.html, published online February 24, 2025
Manuscript received: 2023-09-23

Abstract: This paper proposes a concurrent neural network model to mitigate non-linear distortion in power amplifiers using a basis function generation approach. The model is designed using polynomial expansion and comprises a feedforward neural network (FNN) and a convolutional neural network (CNN). The proposed model takes the basic elements that form the bases as input, defined by the generalized memory polynomial (GMP) and dynamic deviation reduction (DDR) models. The FNN generates the basis function and its output represents the basis values, while the CNN generates weights for the corresponding bases. Through the concurrent training of FNN and CNN, the hidden layer coefficients are updated, and the complex multiplication of their outputs yields the trained in-phase/quadrature (I/Q) signals. The proposed model was trained and tested using 300 MHz and 400 MHz broadband data in an orthogonal frequency division multiplexing (OFDM) communication system. The results show that the model achieves an adjacent channel power ratio (ACPR) of less than −48 dB within a 100 MHz integral bandwidth for both the training and test datasets.

Keywords: basis function generation; digital predistortion; generalized memory polynomial; dynamic deviation reduction; neural network

Citation (Format 1): SHAO J F, HONG X, WANG W J, et al. A basis function generation based digital predistortion concurrent neural network model of RF power amplifier [J]. ZTE Communications, 2025, 23(1): 71 – 77. DOI: 10.12142/ZTECOM.202501009
Citation (Format 2): J. F. Shao, X. Hong, W. J. Wang, et al., "A basis function generation based digital predistortion concurrent neural network model of RF power amplifier," *ZTE Communications*, vol. 23, no. 1, pp. 71 – 77, Mar. 2025. doi: 10.12142/ZTECOM.202501009.

1 Introduction

With the growing demand for high-throughput wireless communications, system bandwidths continue to expand. However the use of orthogonal frequency division multiplexing (OFDM) modulation results in a high peak-to-average power ratio (PAPR)[1]. The nonlinear behavior of power amplifiers (PAs) often leads to compression of high-dynamic-range signals, causing significant signal transmission distortion and upgraded error vector magnitude (EVM) at the receiver, even in scenarios with a high signal-to-noise ratio (SNR)[2]. Therefore, PA behavior modeling and corresponding anti-compression techniques, such as digital predistortion (DPD), play an important role in establishing a robust wireless communication system[3].

The wider bandwidth leads to the existing polynomial expansion models less precise for PA behavior modeling and digital predistortion techniques. Traditional DPD methods, like the generalized memory polynomial (GMP)[4] or dynamic deviation reduction (DDR) model[5], rely on polynomial expansion. However, increasing bandwidth requires higher polynomial orders, which introduces a high correlation among the polynomial's high-order terms, thereby making the traditional models sensitive to noise[6]. Additionally, conventional models require more delay taps and computational resources for high bandwidth signal transmission to radio frequence (RF) PA, complicating their integration with nonlinear bases[7].

Recent research and data analysis indicate that neural networks (NNs) have excellent performance in data feature extraction, data fitting, and model generalization. As a result, the use of NN in DPD has received increased attention and application[8]. For example, a feed-forward NN was proposed in Ref. [9], achieving improvements in both linearity and stability. Similarly, in Refs. [10] and [11], two-stage network models were proposed, achieving good performance metrics such as adjacent channel power ratio (ACPR) and normalized mean square error (NMSE). In Ref. [12], a novel residual NN structure connects residual learning and PA nonlinearity, providing

This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. HC-CN-20220722010.

better performance than conventional models.

Existing methods often achieve better performance by increasing the number of parameters, which in turn significantly raises model complexity. In real-time applications, optimizing model complexity is a crucial aspect of the integration of RF-DPD and NN. To reduce both the training data length and the number of basis functions, Ref. [13] proposed a model that combines an efficient uncorrelated equation selection mechanism with orthogonal least squares. Another model proposed in Ref. [14] is a spare gated dynamic NN DPD model that linearizes the PA for varying transmission configurations, thereby reducing model complexity.

Several new models have been introduced to address the issues of performance and complexity that the classic NN model could not handle. These new models can effectively linearize RF PAs in broadband communications while reducing complexity. For instance, Ref. [15] introduced a novel augmented convolutional NN-based DPD that can linearize concurrent multiband PAs. Additionally, Ref. [16] proposed a novel block-oriented time-delay NN to alleviate the deterioration of linearization performance. Ref. [17] proposed a novel RNN-based behavioral model that reduces complexity and enhances linearization performance by applying the complete phase-gated Just Another Network (JANET) unit. These new models are more suitable for PA-DPD in wide bandwidths and provide better nonlinear modeling capabilities to extract PA features for DPD.

In this paper, we present a DPD concurrent NN model based on an FNN and a convolutional neural network (CNN). The basic inputs of this model are obtained through polynomial expansion of the GMP+DDR model. This concurrent NN model overcomes nonlinear distortions such as amplitude modulation to amplitude modulation (AM-AM) distortion and amplitude modulation to phase modulation (AM-PM) distortion in RF-PAs. Our experimental data comprises OFDM signals with bandwidths of 300 MHz or 400 MHz. Our model aims at reducing non-linear compression to improve the ACPR of the output signal, with a target of at least −48 dB within a 100 MHz integral bandwidth. In addition to its function of basis function generation, the proposed model seeks to have engineering feasibility and low complexity.

2 Mathematical Model of DPD

To enhance the efficiency of PAs, existing methods aim to compress the power regression range as much as possible. The PA's gain does not maintain linearity when the input signal amplitude of the amplifier output section approaches the 1 dB compression point. Typically, power amplifiers feature nonlinear effects such as AM-AM, AM-PM, and time memory. Traditional narrowband amplifiers can be modeled using polynomial expressions, with the Volterra series serving as one of the most representative mathematical models. Eq. (1) describes a P-order and M-length Volterra series.

$$\tilde{y}(n) = \sum_{p=0}^{P}\sum_{m}^{M} h_{p,m} \prod_{l=1}^{p} \tilde{x}(n - m_l) \tag{1},$$

where m denotes the length of the memory effect and p denotes the maximum order of the basis. Similar to the solution of the Wiener filter, the concatenated signal terms in Eq. (1) serve as the bases of polynomial expansion, while the corresponding coefficients $h_{p,m}$ are their respective weights.

While Volterra can effectively describe nonlinear compression with memory effects, the concatenated multiplication of signals introduces a great deal of computational effort and complexity. The memory polynomial (MP) model[17] replaces the concatenated multiplication of signals with a modulus-valued term based on the Volterra series. The MP model can be simplified in the time domain:

$$\tilde{y}(n) = \sum_{p=0}^{P}\sum_{m}^{M} h_{2p,m} \tilde{x}(n - m)|\tilde{x}(n - m)|^{2p} \tag{2}.$$

In the MP model, the basis function becomes the signal multiplied by the signal's ground modulus term. This modification leads to a significant reduction in the computational effort required by the network. However, as the bandwidth increases further, the MP model faces the issue of lower accuracy.

The GMP model extends the composition of the bases based on the MP model. It can describe the nonlinear compression model at larger bandwidths and can be simplified in the time domain as:

$$\tilde{y}(n) = \sum_{p=0}^{P}\sum_{l}^{L}\sum_{m}^{M} h_{2p,l,m} \tilde{x}(n - l)\left|\tilde{x}(n - m)\right|^{2p} \tag{3}.$$

The GMP model extends the influence of time memory effects in the composition of the basis functions, which is relevant to the scenario of wideband communication.

The DDR model[5] is also built on the MP model. However, it differs from the GMP model by placing more emphasis on the aliasing effects of wideband signals. It can be represented in the time domain as:

$$\tilde{y}(n) \approx \sum_{p=0}^{P}\sum_{k}^{K} a_{2p,i,m}\left|x(n)\right|^{2p} x(n - k) + \sum_{p=0}^{P}\sum_{l}^{L} b_{2p-2j,n}\left|x(n)\right|^{2p-2} x^2(n - l)x(n - l) \tag{4}.$$

The DDR model can be divided into two parts. As shown in Eq. (4), the first part is the MP model, while the second part describes the nonlinear compression of the signal after aliasing under the memory effect.

In this paper, since the data we use are wideband signals and the main requirement for the proposed model is better performance, we use the GMP+DDR model as the reference mathematical model for this paper. It can be written as:

$$\tilde{y}(n) \approx \sum_{p=0}^{P}\sum_{i=0}^{I}\sum_{k}^{K} a_{2p,i,m}\left|x(n-i)\right|^{2p} x(n-k) + \sum_{p=0}^{P}\sum_{j=1}^{J}\sum_{l}^{L} b_{2p-2j,n}\left|x(n-j)\right|^{2p-2} x^2(n-l)x(n-l) \quad (5).$$

Eq. (5) combines the features of the GMP and DDR models. Both the memory effect and aliasing of broadband signals are covered to ensure that the model can achieve optimal performance. The dataset composition will also refer to the mathematical model shown in Eq. (5).

3 Designing of DPD NN Model

3.1 Basis Function Generation and Recognition

To explain the generation of basis functions, we first clarify the input dataset structure. Our basis function formulation, based on the GMP+DDR model for wideband applications, draws inspiration from the methodologies presented in Refs. [4] and [5]. The intermodulation terms in Eq. (5) are highly suitable for modeling the wideband PA. Therefore, we establish the dataset format based on the fundamental components in Eq. (6).

$$\begin{aligned}&\tilde{\boldsymbol{x}}(n) = [Re(x(n-12)),\cdots,Re(x(n)),\cdots,Re(x(n+11)),\\&Im(x(n-12)),\cdots,Im(x(n)),\cdots,Re(x^2(n)),\cdots,Im(x^2(n)),\cdots,\\&|x(n)|^2,\cdots,|x(n)|^4,\cdots,|x(n)|^6,\cdots]^T,\\&\boldsymbol{X} = [\cdots\tilde{\boldsymbol{x}}(n-1),\tilde{\boldsymbol{x}}(n),\tilde{\boldsymbol{x}}(n+1),\cdots] \quad (6).\end{aligned}$$

The dataset consists of multiple vectors, as shown in Eq. (6), indicating the input terms and memory depth. The input elements include the signal, the square of the signal, and the even-square term of the signal's modulus. And they all stem from the GMP+DDR model, as detailed in Eq. (5). The model has an order of 7 and a memory depth of 24, spanning from −12 to 11. Since the neural network library we use (PyTorch) is less compatible with complex numbers, the proposed model is trained using real-valued data. For this purpose, the real and imaginary parts of the signal are split and used to construct the dataset.

To linearize the bases of individual nonlinear terms, we propose the use of fully connected (FC) layers to combine all elements. This approach enables the number of basis elements to be established by the number of neurons within the FC layer. The output of each layer is then nonlinearly activated to generate a nonlinear basis. Additionally, based on the survey results, at least three FC layers are sufficient to produce the majority of nonlinear combinations. Then the outputs would be activated by the nonlinear function to ensure their nonlinearity.

In this paper, the basis generation network (BGN) based on an FNN is illustrated in Fig. 1. The weight matrix within the FC layer adjusts the coefficients of the input terms, which are optimized through training feedback. As shown in Fig. 1, the length of the FC layers decreases in the forward direction of the arrays. Therefore, the number of neurons and the output of each FC layer are decreased. It is similar to FNN selecting bases for each hidden layer. Regarding the activation function, Rectified Linear Unit (ReLU) leads to faster loss convergence compared to other activation functions based on test results. This improvement can be attributed to ReLU's superior sparsity. Consequently, each FC layer's output in the basis function generation model is activated by ReLU. $B_{i/q-1}$ and $B_{i/q-N}$ in Fig. 1 denote the real or imaginary parts of the first and N-th substrates generated, respectively. Since the DPD model is a real-valued training NN model, the BGN has two identical structures as shown in Fig. 1. The notation "i/q" represents in-phase or quadrature components, while "-1" or "$-N$" serves as a label for the bases. These labels have no real physical meaning and are solely used to distinguish the bases and correspond to the weights.

3.2 Structure of Concurrent NN Model

Fig. 2 depicts the proposed concurrent neural network model, comprising an FNN and a CNN.

The left side of Fig. 2 displays the FNN model utilized to generate basis functions, as described in Section 3.1. Since the proposed model is trained using real numbers, the FNN-based function generation model has two sets of three FC layers. The basis generation function only relies on the coefficients of each hidden layer in the FNN model. On the right

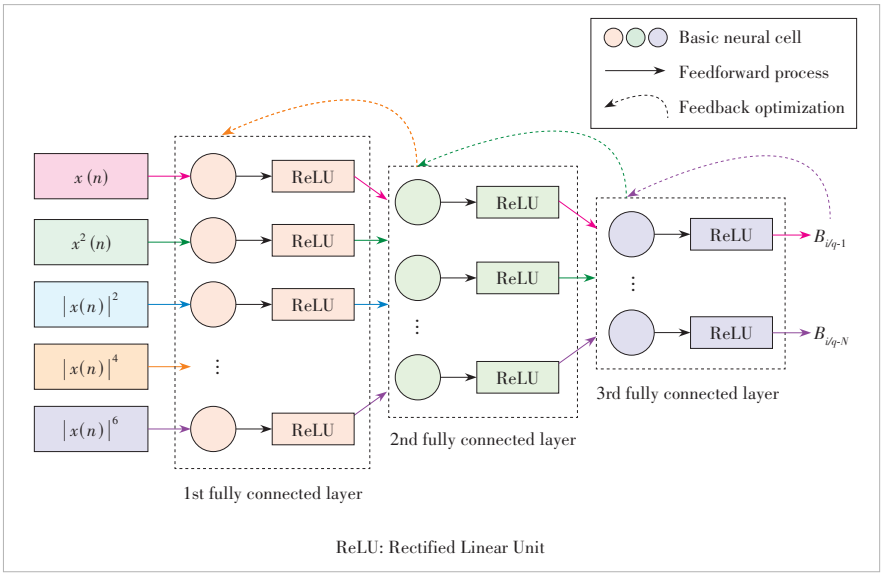


Figure 1. Proposed basis generation function

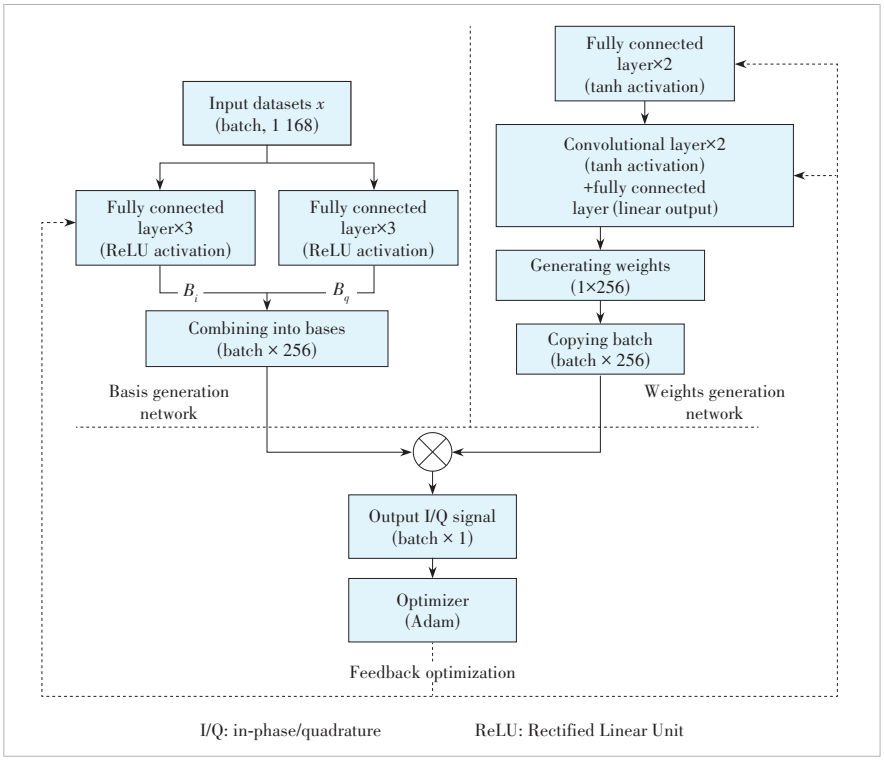


Figure 2. Proposed parallel concurrent neural network for digital predistortion

side of Fig. 2, the weights generation network (WGN) based on a CNN is illustrated. It contains three fully connected layers and two convolutional layers, as shown in Fig. 2. To ensure that the weights generation does not change by the input, the input to the WGN is fixed at a constant value (set to 1 in the following training). The input to the CNN model is derived from the output of the two FC layers positioned above it. Therefore, the input of the WGN has no real physical meaning. WGN aims to help us find a proper weight through a training process as the loss decreases. Similar to the bases, we assume that when the loss converges to a very low level, the weights will be optimized to best describe the nonlinear compression features in the trained data. Subsequently, the outputs of the FNN and CNN are trained through the projection layer at the bottom of Fig. 2 to generate the in-phase/quadrature (I/Q) data. The projection layer is tailored to perform complex multiplication accurately by incorporating appropriate dimensional changes, aligning the I/Q data, and producing the final output. The mean squared error (MSE) of the model's output is computed using the validation set (valset) as the model loss. This calculated model loss is then utilized as feedback to fine-tune the coefficients in all hidden layers of the proposed model. Furthermore, the weights generated by the CNN model are insensitive to the data fed into the model. Therefore, if the model is executed on a hardware platform such as a Field-Programmable Gate Array (FPGA), only the FNN network needs to be deployed. The trained FNN network carries out only linear operations and is easily implemented in engineering applications.

In essence, the proposed FNN-CNN concurrent model achieves DPD through time-domain fitting. The FNN model generates the basis functions based on the GMP+DDR model via model training, while the CNN model produces the weights using the coefficients from each hidden layer. Both models are jointly optimized to minimize the loss. The final output is obtained by multiplying the basis functions with the weights, after which the valset is used to compute the loss. A model with such a structure, concurrently trained by CNN and FNN, is dubbed a concurrent neural network model.

4 Training Process and Results

4.1 Dataset Use Cases

Two datasets from different RFs are available in the OFDM communication system, with bandwidths of 300 MHz and 400 MHz. Each dataset has 10×16 384 samples. We use eight of ten feedback signals as training sets (trainset), the other two of them as test sets (testset), and the corresponding transmission signal fed into the PA as the validation set.

Fig. 3 illustrates the nonlinear compression of the datasets containing both 300 MHz and 400 MHz bandwidths. The PA significantly compresses the signal amplitude. As shown in Table 1, the compressed signal produces severe out-of-band leakage and nonlinear distortion. Table 1 presents the frequency-domain performance of the datasets, and the ACPR is calculated at integral bandwidths of 100 MHz and 20 MHz. The primary goal of this paper is to minimize the ACPR (with a target of at least −48 dB) of the output generated by the proposed model by employing optimization and training techniques.

The proposed model, as outlined in Section 3, aims to eliminate the out-of-band nonlinear distortion through time-domain fitting of the trainset to the valset. Moreover, we evaluate the model's effectiveness through the ACPR at an integrated bandwidth of 100 MHz.

4.2 Training Results

The datasets shown in Section 4.1 are utilized to train the proposed model with the MSE serving as the loss function. Eight out of ten training sets are selected randomly for the training process, and the remaining two datasets are used as testsets to evaluate the trained model's performance.

Fig. 4 depicts the evolution of the model output's MSE over 20 000 training epochs. The blue curve represents the

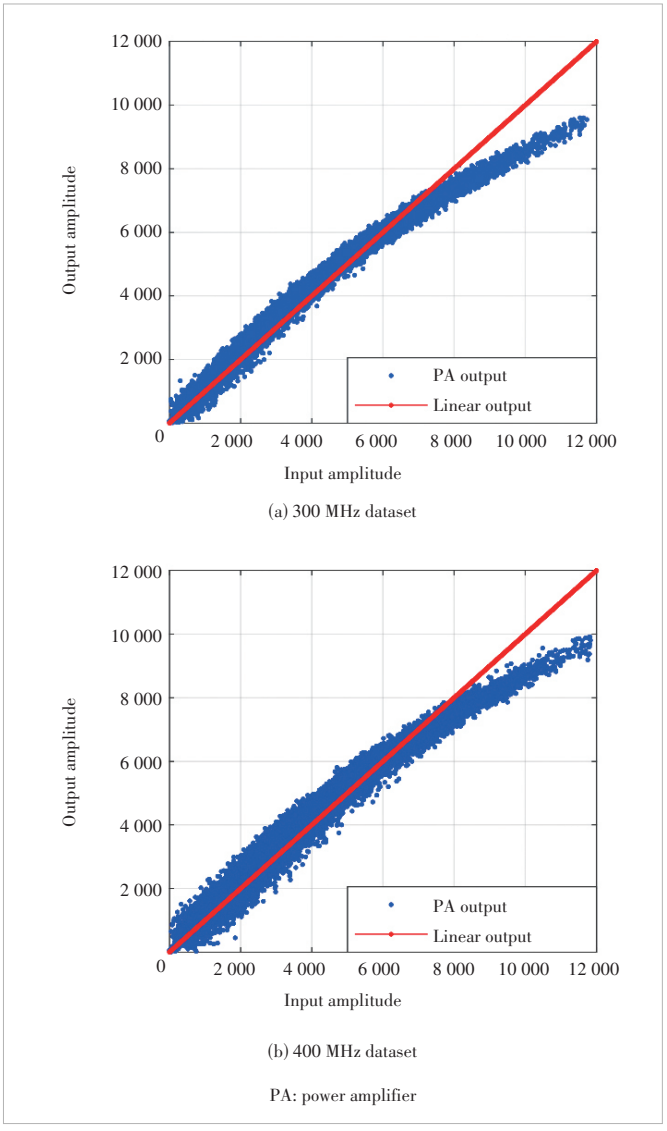


Figure 3. Demonstration of training and validation sets

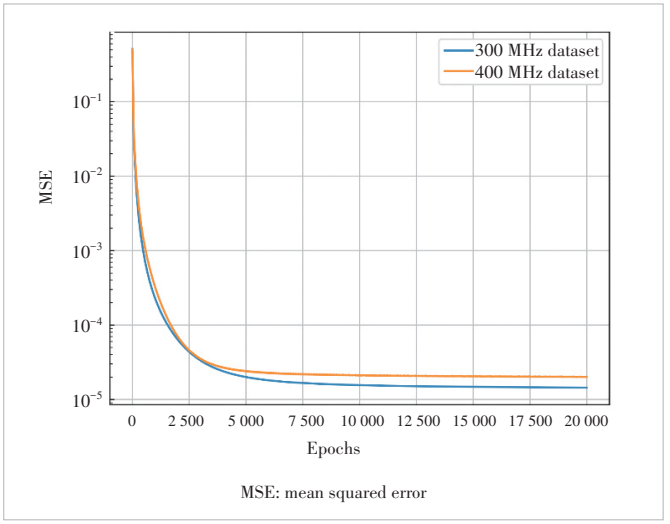


Figure 4. Model training loss

300 MHz dataset, whose MSE drops sharply within the first 5 000 epochs, reaching a plateau thereafter and converging below 2×10^{-5} around 12 500 epochs. The final loss of the 300 MHz dataset after 20 000 epochs is 1.4×10^{-5}. The orange curve, representing the 400 MHz dataset, exhibits a similar downward trajectory, albeit with a poorer result than the blue curve. Its final loss after 20 000 epochs is 1.97×10^{-5}. Notably, the increase in bandwidth from 300 MHz to 400 MHz does not interfere with the convergence speed. The final convergence value is affected not only by model training but also by the differences in the datasets. Fig. 4 provides evidence that the proposed model performs well across various bandwidths, thereby highlighting its generalizability.

Fig. 5 shows the DPD results of the proposed model and a comparison with the existing Generalized Memory Polynomial (GMP) algorithm. Specifically, the out-of-band power of the proposed model (the blue curve in the figure) is approximately 40 dB lower than that of the feedback signal (the red curve) in Figs. 5a and 5b. In addition, the signal portion of the model output closely matches the signal component of the source signal (the cyan curve) for both 300 MHz and 400 MHz cases. This outcome indicates the remarkable ability of the proposed model to suppress out-of-band nonlinear distortion without compromising the fitting of the signal portion. When compared with the existing algorithms, the proposed model has obvious advantages in suppressing out-of-band leakage. There is nearly 20 dB optimization compared to the GMP model (the

Table 1. Frequency domain performance of datasets

Datasets	ACPR of Left Band/dB		ACPR of Right Band/dB		NMSE
	100 MHz Integral Bandwidth	20 MHz Integral Bandwidth	100 MHz Integral Bandwidth	20 MHz Integral Bandwidth	
300 MHz valset	−54.17	−46.18	−58.04	−48.65	
300 MHz datasets	Max=−21.45 Min=−21.49 Average=−21.47	Max=−21.00 Min=−21.02 Average=−21.00	Max=−21.94 Min=−21.98 Average=−21.96	Max=−21.01 Min=−21.01 Average=−21.01	Max=−16.40 Min=−16.41 Average=−16.40
400 MHz valset	−54.90	−43.29	−53.61	−46.86	
400 MHz datasets	Max=−21.36 Min=−21.40 Average=−21.38	Max=−19.39 Min=−19.48 Average=−9.45	Max=−22.61 Min=−22.88 Average=−22.66	Max=−20.13 Min=−20.22 Average=−20.19	Max=−17.02 Min=−17.05 Average=−17.04

ACPR: adjacent channel power ratio NMSE: normalized mean square error

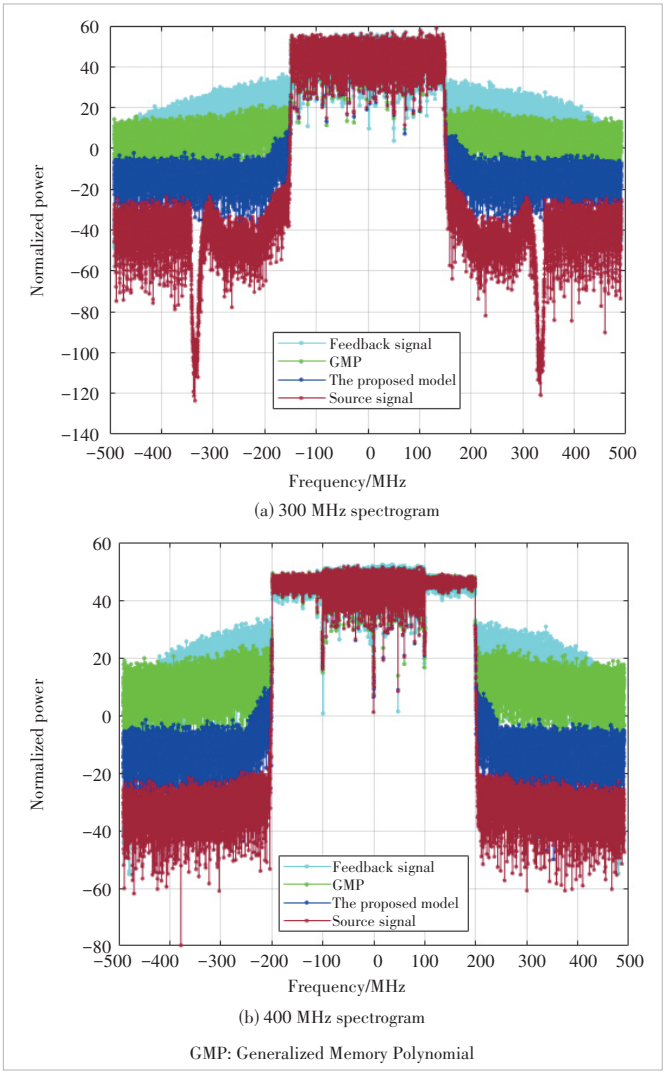


Figure 5. Digital predistortion results of the proposed model

green curve). However, the training results for the 300 MHz and 400 MHz datasets in Fig. 5 still exhibit some out-of-band non-linear distortion, approximately 20 dB higher than the out-of-band part of the validation sets.

Table 2 presents the training and testing results of the proposed model for both the 300 MHz and 400 MHz datasets. The ACPR of both the left and right frequency bands can exceed −48 dB for an integral bandwidth of 100 MHz. Additionally, the model output's NMSE indicates an improvement of nearly −30 dB compared to the initial NMSE of the trainsets, demonstrating that the model output fits the signal portion well. The ACPR of the testsets is also greater than −48 dB; however, it is approximately 1 dB worse than the ACPR of the training sets, and the NMSE training results show similar results.

In addition to evaluating the ACPR of the 100 MHz integral bandwidth, this study also calculates the ACPR of the 20 MHz integral bandwidth to identify why the ACPR produces suboptimal outcomes. By comparing Table 2 with Table 1, it is observable that the performance difference between the model output and the trainsets or testsets remains the same for both 100 MHz and 20 MHz integral bandwidths. Consequently, the expansion of 80 MHz to the periphery does not affect the ACPR results. Instead, the primary factors affecting the ACPR assessment are concentrated within the 20 MHz band boundary.

5 Conclusions

This paper presents a concurrent NN model of RF PA designed to accomplish DPD functions. The proposed model employs the enhanced DDR (GMP+DDR) model as input, which is more suited for modeling the behavior of broadband communication systems. The FNN generates the basis functions, while the CNN generates the weights, with the entire model trained to simultaneously generate their respective optimized values. This study employed eight sets of 300 MHz and 400 MHz data for 20 000 epochs and tested the model with two sets of data. After training and testing, the desired goal of achieving a −48 dB ACPR by 100 MHz integral bandwidth was met for both the trainsets and testsets. The spectrogram shows that the proposed model has a great advantage over the existing algorithms in wider bandwidth scenarios. Moreover, the ACPR was evaluated at 20 MHz integral bandwidth, revealing that the roll-off is the primary limitation of

Table 2. Frequency domain performance of model output

Datasets	ACPR of Left Band/dB		ACPR of Right Band/dB		NMSE
	100 MHz Integral Bandwidth	20 MHz Integral Bandwidth	100 MHz Integral Bandwidth	20 MHz Integral Bandwidth	
300 MHz trainsets	Max=−50.27 Min=−50.77 Average=−50.48	Max=−42.53 Min=−43.15 Average=−42.94	Max=−51.29 Min=−51.99 Average=−51.70	Max=−44.25 Min=−45.14 Average=−44.73	Max=−43.62 Min=−46.12 Average=−46.69
300 MHz testsets	−49.65, −49.48	−42.77, −42.57	−50.71, −50.69	−44.35, −44.25	−44.77, −44.79
400 MHz trainsets	Max=−48.93 Min=−49.36 Average=−49.17	Max=−40.99 Min=−41.49 Average=−41.24	Max=−48.66 Min=−49.10 Average=−48.82	Max=−42.45 Min=−43.40 Average=−43.10	Max=−43.86 Min=−44.31 Average=−44.15
400 MHz testsets	−48.02, −48.17	−40.50, −40.64	−48.13, −48.12	−42.44, −42.54	−42.92, −42.85

ACPR: adjacent channel power ratio NMSE: normalized mean square error

ACPR. This finding can guide future efforts to optimize the proposed model.

References

[1] LIU Z J, HU X, WANG W D, et al. A joint PAPR reduction and digital predistortion based on real-valued neural networks for OFDM systems [J]. IEEE transactions on broadcasting, 2022, 68(1): 223 – 231. DOI: 10.1109/TBC.2021.3132158
[2] LIU Z J, HU X, XU L X, et al. Low computational complexity digital predistortion based on convolutional neural network for wideband power amplifiers [J]. IEEE transactions on circuits and systems II: express briefs, 2022, 69(3): 1702 – 1706. DOI: 10.1109/TCSII.2021.3109973
[3] LIU X, CHEN W H, WANG D H, et al. Robust digital predistortion for LTE/5G power amplifiers utilizing negative feedback iteration [J]. ZTE communications, 2020, 18(3): 49 – 56. DOI: 10.12142/ZTECOM .202003008
[4] LIU Y J, ZHOU J, CHEN W H, et al. A robust augmented complexity-reduced generalized memory polynomial for wideband RF power amplifiers [J]. IEEE transactions on industrial electronics, 2014, 61(5): 2389 – 2401. DOI: 10.1109/TIE.2013.2270217
[5] ZHU A D, DRAXLER P J, YAN J J, et al. Open-loop digital predistorter for RF power amplifiers using dynamic deviation reduction-based Volterra series [J]. IEEE transactions on microwave theory and techniques, 2008, 56(7): 1524 – 1534. DOI: 10.1109/TMTT.2008.925211
[6] BRAITHWAITE R N. Adaptation of a digitally predistorted RF amplifier using selective sampling [J]. ZTE communications, 2011, 9(3): 3 – 12
[7] HU X, LIU Z J, WANG W D, et al. Low-feedback sampling rate digital predistortion using deep neural network for wideband wireless transmitters [J]. IEEE transactions on communications, 2020, 68(4): 2621 – 2633. DOI: 10.1109/TCOMM.2020.2966718
[8] ROSOŁOWSKI D W, JĘDRZEJEWSKI K. Experimental evaluation of PA digital predistortion based on simple feedforward neural network [C]//Proc. 23rd International Microwave and Radar Conference (MIKON). IEEE, 2020: 293 – 296. DOI: 10.23919/ MIKON48703.2020.9253814
[9] FENG X, FEUVRIE B, DESCAMPS A S, et al. Digital predistortion method combining memory polynomial and feed-forward neural network [J]. Electronics letters, 2015, 51(12): 943 – 945. DOI: 10.1049/el.2015.0276
[10] WU H B, CHEN W H, LIU X, et al. A uniform neural network digital predistortion model of RF power amplifiers for scalable applications [J]. IEEE transactions on microwave theory and techniques, 2022, 70(11): 4885 – 4899. DOI: 10.1109/TMTT.2022.3205930
[11] JUNG S, KIM Y, WOO Y, et al. A two-step approach for DLA-based digital predistortion using an integrated neural network [J]. Signal processing, 2020, 177: 107736. DOI: 10.1016/j.sigpro.2020.107736
[12] WU Y B, GUSTAVSSON U, AMAT A G I, et al. Residual neural networks for digital predistortion [C]//Proc. 2020 IEEE Global Communications Conference. IEEE, 2020: 1-6. DOI: 10.1109/globecom42002.2020.9322327
[13] LÓPEZ-BUENO D, MONTORO G, GILABERT P L. Training data selection and dimensionality reduction for polynomial and artificial neural network MIMO adaptive digital predistortion [J]. IEEE transactions on microwave theory and techniques, 2022, 70(11): 4940 – 4954. DOI: 10.1109/TMTT.2022.3209214
[14] JIANG C Y, YANG G C, HAN R L, et al. Gated dynamic neural network model for digital predistortion of RF power amplifiers with varying transmission configurations [J]. IEEE transactions on microwave theory and techniques, 2023, 71(8): 3605 – 3616. DOI: 10.1109/TMTT.2023.3241612
[15] JARAUT P, ABDELHAFIZ A, CHENINI H, et al. Augmented convolutional neural network for behavioral modeling and digital predistortion of concurrent multiband power amplifiers [J]. IEEE transactions on microwave theory and techniques, 2021, 69(9): 4142 – 4156. DOI: 10.1109/TMTT.2021.3075689
[16] JIANG C Y, LI H M, QIAO W, et al. Block-oriented time-delay neural network behavioral model for digital predistortion of RF power amplifiers [J]. IEEE transactions on microwave theory and techniques, 2022, 70(3): 1461 – 1473. DOI: 10.1109/TMTT.2021.3124211
[17] KOBAL T, LI Y, WANG X Y, et al. Digital predistortion of RF power amplifiers with phase-gated recurrent neural networks [J]. IEEE transactions on microwave theory and techniques, 2022, 70(6): 3291 – 3299. DOI: 10.1109/TMTT.2022.3161024
[18] MONDAL R, RISTANIEMI T, DOULA M. Genetic algorithm optimized memory polynomial digital pre-distorter for RF power amplifiers[C]//Proc. International Conference on Wireless Communications and Signal Processing. IEEE, 2013: 1 – 5. DOI: 10.1109/WCSP.2013.6677117

Biographies

SHAO Jianfeng (sjf1996717@stu.xjtu.edu.cn) received his BS degree in electronic information engineering from the School of Information Engineering, North China University of Water Resources and Electric Power in 2018, MS degree in circuits and systems from the School of Physics and Electronic Engineering, Ningxia University, China in 2021. He is pursuing a PhD degree in information and communication engineering at Xi'an Jiaotong University, China. His research interests include array signal processing and digital pre-distortion.

HONG Xi received his BS degree in information engineering from the School of Electronics and Information Engineering, Xi'an Jiaotong University, China in 2012, and PhD degree in information and communication engineering from the School of Information and Communication Engineering, Xi'an Jiaotong University in 2022. He is currently an engineer at the School of Information and Communication Engineering, Xi'an Jiaotong University. His research interests include array signal processing, signal processing in communication systems, multipath mitigation in navigation, and GNSS physical layer interference detection.

WANG Wenjie received his BS, MS, and PhD degrees in information and communication engineering from Xi'an Jiaotong University, China in 1993, 1998, and 2001, respectively, where he is currently a professor. His main research interests include information theory, broadband wireless communications, signal processing in communication systems, and array signal processing.

LIN Zeyu received his BS and MS degrees at the School of Information and Communication Engineering, Xi'an Jiaotong University, China in 2018 and 2021, respectively. He is currently a senior RF algorithm architect in the RHP department of ZTE Corporation. His research interests include PA nonlinear system behavioral modeling and RF systems linearization.

LI Yunhua received his BE degree in communications engineering from Shandong University (Weihai), China in 2012 and PhD degree in military communications from the School of Telecommunications Engineering, Xidian University, China in 2017. He is currently a senior engineer of RF algorithms in ZTE Corporation. His research interests include communication signal processing, wireless communications, digital predistortion modeling of nonlinear systems, and more.

A Wide Passband Frequency Selective Surface with Angular Stability

TANG Xingyang[1], SUI Jia[1], FU Jiahui[1], YANG Kaiwen[2], ZHAO Zhipeng[2]
(1. Harbin Institute of Technology, Harbin 150001, China;
2. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202501010
https://kns.cnki.net/kcms/detail/34.1294.TN.20250225.1143.002.html, published online February 26, 2025
Manuscript received: 2023-09-24

Abstract: A wide passband frequency selective surface (FSS) is proposed using a five-layer stacked structure. The proposed structure applies four layers of dielectric plates and five layers of metal patches to provide a passband and exhibits more stable frequency responses and lower insertion loss under wide-angle oblique incidence compared with the typical three-layer metal-dielectric structure. According to the simulation results, the proposed FSS can achieve a passband range of 1.7–2.7 GHz with an insertion loss of less than 0.5 dB and a relative bandwidth of 44.1%, and it can preserve stable transmission characteristics with the incident angle ranging from 0° to 45°.

Keywords: frequency selective surface (FSS); wide bandwidth; low insertion loss

Citation (Format 1): TANG X Y, SUI J, FU J H, et al. A wide passband frequency selective surface with angular stability [J]. ZTE Communications, 2025, 23(1): 78 – 84. DOI: 10.12142/ZTECOM.202501010
Citation (Format 2): X. Y. Tang, J. Sui, J. H. Fu, et al., "A wide passband frequency selective surface with angular stability," *ZTE Communications*, vol. 23, no. 1, pp. 78 – 84, Mar. 2025. doi: 10.12142/ZTECOM.202501010.

1 Introduction

Frequency selective surfaces (FSS) are a periodic array structure composed of metal patch units on the dielectric substrate or aperture elements on the metal screen. FSS arrays are essentially spatial filters that can select the working frequency and polarization mode of electromagnetic waves, such as transverse electric (TE) and transverse magnetic (TM), according to the relationship between electric and magnetic fields and the incident plane[1–3]. The patch type FSS shows the band-stop characteristic and the aperture type FSS shows the band-pass characteristic. FSS is frequently employed in radomes, antenna reflectors, electromagnetic shielding, etc[4]. For the radome loaded on the filter antenna, the wave transmission characteristic is mainly determined by the loaded FSS array. FSS is a versatile structure that plays a crucial role in controlling and manipulating electromagnetic waves for various applications, with their characteristics determined by the design of the FSS unit, arrangement period, and dielectric properties of the substrate.

Broadband communication systems have proposed stricter bandwidth requirements in recent years[5]. The wide passband FSS can achieve low insertion loss electromagnetic wave transmission under a broad band and large angle incidence. Multiple works have studied the design of wide passband FSS. In Refs. [6] and [7], a planar broadband FSS composed of three layers of patches is introduced. It exhibits a relative bandwidth of 42% under vertical incidence, although its insertion loss deteriorates significantly at large angles of incidence. A capped dielectric inserted perforated metallic plate bandpass frequency selective surface is reported in Ref. [8]. It can achieve 40° oblique incident stability with a low profile, but its performance is susceptible to fabrication tolerances. In Ref. [9], a three-dimensional FSS with sharp roll-off sidebands is proposed, which has 62% relative bandwidth and sharp roll-off sidebands under the incident wave of TE polarization modes. The demands of dual polarization applications cannot be met by this 3D FSS since it only supports a single polarization wave. A dual-band FSS alternative solution with a complex manufacturing process and a high-dimensional structure is provided in Ref. [10] and can satisfy the demands of applications involving curved surfaces[11]. Ref. [12] proposes a broadband FSS load with charged inductance. Based on multilayer cascaded FSS, the characteristics of broadband, low profile, and miniaturization are achieved by increasing the lumped inductance. A passband with a reflection coefficient below −10 dB was obtained at 0.1 – 1.2 GHz, and a stopband with a transmission coefficient below −10 dB was obtained at 5.8 – 12 GHz. Ref. [13] proposes a relatively simple FSS structure for antenna beam control applications, where the FSS

This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. IA20220800001.

structure utilizes a small number of active components to achieve reconfigurability, good transmission and reflection characteristics required for wireless communication applications. In Ref. [14], the FSS bending effect is considered and the performance of FSS deteriorates with increasing curvature. Few studies have been conducted on broadband oblique-incidence stable FSS, which cannot satisfy communication systems' demands for broadband stable FSS. The goal of this paper is to investigate this issue and propose a better framework.

In this paper, a bandpass FSS with a patch-dielectric cascading structure is proposed. It uses two kinds of circular patches with different radii and cross-slotted patches to achieve three transmission poles in the passband. Simulation results indicate a satisfactory −0.5 dB bandwidth spanning from 1.7 GHz to 2.72 GHz, maintaining angular stability from 0° to 45°. Compared with the three-layer FSS, the designed FSS has a wider passband and better oblique incidence frequency response. The relative bandwidth of this FSS is 45.2%, which has low insertion loss and oblique incident stability under both TE and TM polarization. In the meantime, equivalent circuits are provided to check the accuracy of electromagnetic simulations.

2 Design Principle

The design principle of the angular stability FSS is essentially electromagnetic wave impedance matching under oblique incidences. An electromagnetic wave vector is only along its propagation direction. Assume that the vectors of incident, reflected and transmitted waves are $\boldsymbol{k}_i$, $\boldsymbol{k}_r$ and $\boldsymbol{k}_t$, respectively. The incident, reflection and refraction angles are θ_i, θ_r and θ_t.

The phase matching condition on the interface is $\boldsymbol{k}_i \sin\theta_i = \boldsymbol{k}_r \sin\theta_r = \boldsymbol{k}_t \sin\theta_t$, where $\boldsymbol{k}_i = \boldsymbol{k}_r = \boldsymbol{k}_1$, and $\boldsymbol{k}_t = \boldsymbol{k}_2$. From the above formula, the Snell's law can be obtained. There are two laws in total, namely, the Snell reflection law and the Snell refraction law. The reflection law states that the reflection angle is equal to the incident angle, that is, $\theta_i = \theta_r$. The law of refraction is expressed as the relationship between the refractive angle and the incident angle, that is, $\sin\theta_i/\sin\theta_t = \boldsymbol{k}_2/\boldsymbol{k}_1$, where $\boldsymbol{k}_1 = \omega\sqrt{\varepsilon_1\mu_1}$ and $\boldsymbol{k}_2 = \omega\sqrt{\varepsilon_2\mu_2}$. The Snell's laws reflect the laws of reflection and refraction of electromagnetic waves and have a wide range of applications.

According to the boundary conditions, the polarization characteristics will not change when waves are reflected and refracted on the plane boundary, regardless of whether it is a vertically polarized plane wave or a parallel polarized plane wave. When the electromagnetic wave is oblique incidence, the reflection and transmission coefficients are related to the polarization characteristics of the wave. It is viable to derive the formula for the plane wave's reflection and transmission coefficients with two different polarization characteristics.

$$\Gamma_{\perp} = \frac{\eta_2\cos\theta_i - \eta_1\cos\theta_t}{\eta_2\cos\theta_i + \eta_1\cos\theta_t} \tag{1},$$

$$\tau_{\perp} = \frac{2\eta_2\cos\theta_i}{\eta_2\cos\theta_i + \eta_1\cos\theta_t} \tag{2},$$

where η_1 and η_2 are the characteristic impedance of medium1 and medium2, respectively. Similarly, for parallel polarized plane waves, we obtain that

$$\Gamma_{\parallel} = \frac{\eta_1\cos\theta_i - \eta_2\cos\theta_t}{\eta_1\cos\theta_i + \eta_2\cos\theta_t} \tag{3},$$

$$\tau_{\parallel} = \frac{2\eta_2\cos\theta_i}{\eta_1\cos\theta_i + \eta_2\cos\theta_t} \tag{4}.$$

It is well known that when the electromagnetic wave is oblique incidence on the surface of the medium, partial reflection and partial transmission occur. In particular, Eqs. (1) and (3) demonstrate that, in the case of vertical polarization, when the characteristic impedance of the two media is $\eta_1/\eta_2 = \cos\theta_i/\cos\theta_t$, the reflection coefficient is equal to 0; in the case of horizontal polarization, when the characteristic impedance of the two media satisfies $\eta_1/\eta_2 = \cos\theta_i/\cos\theta_t$, the reflection coefficient is equal to 0, and there is no reflected wave. The key to improving the stability of oblique incidence is the matching of wave impedance under oblique incidence.

The wave vector of an electromagnetic wave and the normal vector at the interface of the incident medium form the incident plane. There are two different types of polarized waves in the case of oblique incidence: TE polarization (where the electric field is parallel to the incident plane) and TM polarization (where the magnetic field is parallel to the incident plane). The vector transmission line equation can be used to determine the characteristic impedance of the transmission line for the two polarization modes or the impedance of free space waves for the two polarization modes with oblique incidence.

$$Z^{\mathrm{TE}} = \sqrt{\frac{\mu}{\varepsilon}}\frac{1}{\sqrt{1 - \dfrac{1}{\omega^2\varepsilon_t\mu_n}}} \tag{5},$$

$$Z^{\mathrm{TM}} = \sqrt{\frac{\mu}{\varepsilon}}\sqrt{1 - \frac{k_t^2}{\omega^2\varepsilon_n\mu_t}} \tag{6}.$$

By substituting the relevant formula and the permeability and permittivity of free space, the characteristic impedance of free space under vertical incidence is 377 Ω, and the above formula is simplified to $Z_0^{\mathrm{TM}} = 377\cos\theta$, $Z_0^{\mathrm{TE}} = 377/\cos\theta$. The wave vector of electromagnetic waves and the normal vector at

the interface of the incident medium form the incident plane. When the electric field is perpendicular to the incident plane, it is a TE wave; when the magnetic field is perpendicular to the incident plane, it is a TM wave. Fig. 1 shows that the wave impedance of free space under oblique incidence is related to the polarization mode of electromagnetic waves, and the wave impedance of free space under TE and TM modes has the opposite trend with the incident angle.

3 Configuration and Discussion

3.1 FSS Configuration

The multi-layer FSS performs better than the single-layer FSS in terms of bandwidth, oblique incidence transmission coefficients, passband flatness and other factors. In this paper, a multi-layer metal-dielectric stack structure is used to design a broadband oblique-incidence stable FSS. The top and bottom layers feature circular patches, while the middle layer employs cross-shaped slots to provide the first resonance point and control the coupling between the upper and lower layers to generate a second resonance point. The circular patches reduce the effective size variations for oblique incidence. The cross-shaped slots can minimize changes in the electric field direction. To further increase the number of resonance points, an additional layer of dielectric and patch structures is introduced, resulting in a four-layer dielectric and five-layer patch FSS unit.

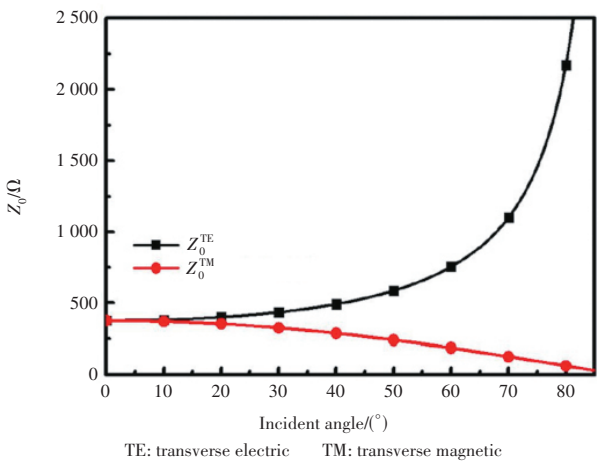


Figure 1. Wave impedance of free space under oblique incidence

As shown in Fig. 2, based on the three-layer patch FSS, a five-layer metal patch structure is proposed to expand the bandwidth and reduce the insertion loss deterioration under oblique incidence. The relative permittivity of each dielectric substrate in designed FSS is 3, and the loss tangent is 0.001 3 @ 10 GHz. This five-layer design has been determined to be a simple and efficient FSS structure through contrasting various patch and analysis structures.

Fig. 2a shows the three-layer FSS, which consists of two layers of circular patches with the same radius and a cross slot in the middle layer to form a resonant structure. This structure typically has two transmission poles available. When the upper patch completely covers the cross slot, the cross slot determines the position of the low-frequency transmission pole, the thickness of the medium controls the distance between the two poles, and the radius of the patch controls the passband frequency. Fig. 2b shows the proposed FSS structure, adding two layers of circular patches with a larger radius to the traditional two-layer circular patch structure to increase the transmission poles and improve the oblique incidence stability. Fig. 2c shows the dimensions of the cross slot in the middle layer.

To achieve better results, we discuss and optimize the main parameters of the structure. Fig. 3a shows the simulation results of the frequency responses with changing length l of the cross slot. It is shown that the length of the cross slot has more effect on the first and second resonant frequencies, and less effect on the third resonant frequency. As l increases from 23.14 mm to 25.14 mm, the first and second transmission poles gradually move away from each other, and the relative bandwidth increases from 42.4% to 47.8%. Similarly, Fig. 3b shows the simulation results of the frequency responses for changing the top patch radius R_1. Changing R_1 has little effect on the first transmission pole, but significantly affects the second and third poles. When R_1 increases to 15.7 mm, FSS can no longer support a passband.

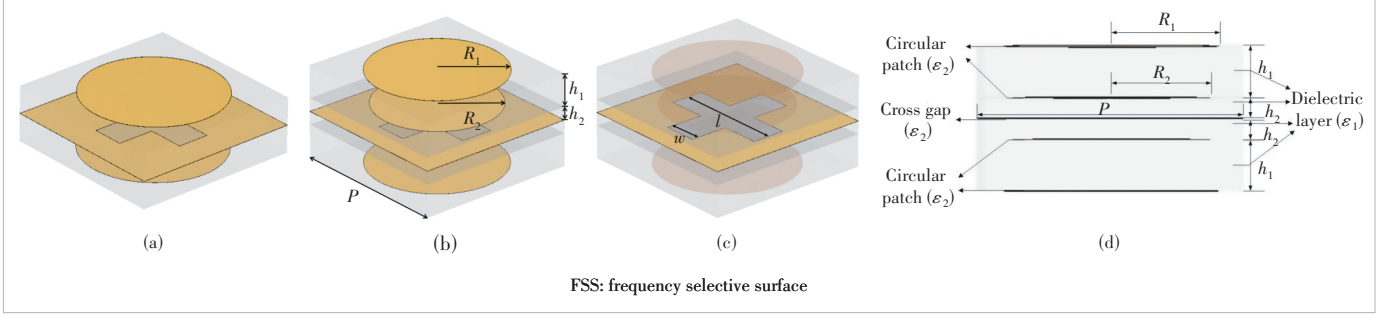


Figure 2. Geometry of the proposed wideband FSS: (a) 3D view of three-layer FSS; (b) 3D view of the proposed FSS (unit: mm; P = 36, R_1 = 14.7, R_2 = 13.5, h_1 = 7, and h_2 = 2.81); (c) 3D view of the interlayer cross slot of the proposed FSS (unit: mm; l = 24.14 and w = 7.29); (d) side view

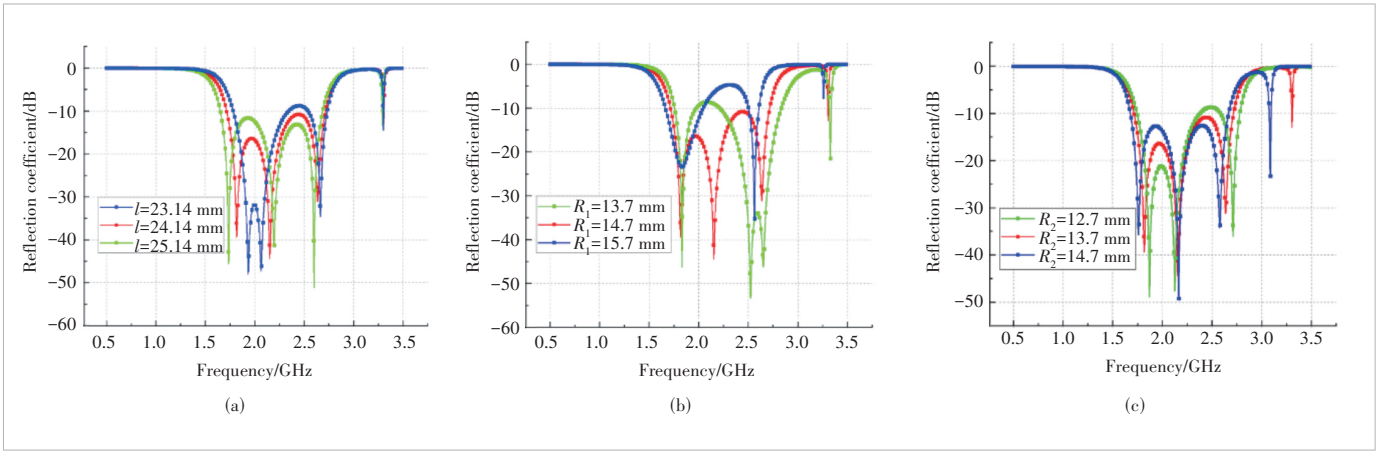


Figure 3. Simulation scattering parameter results of FSS unit cell: (a) results with different cross slot lengths *l*; (b) results with different top patch radii R_1; (c) results with different middle patch radii R_2

Fig. 3c shows the simulation results of the frequency responses with different middle patch radii R_2. The value of radius R_2 has less effect on the second transmission pole but more effect on the first and third transmission poles. As R_2 increases, the first and third transmission poles move to lower frequencies, and the relative bandwidth gradually narrows. In addition, the frequency response of FSS is very sensitive to the thickness of the dielectric substrate. Different thicknesses will affect the position of the resonance frequency, bandwidth, oblique incidence performance, etc., so it needs to be considered comprehensively in the design.

3.2 Simulation Results and Discussion

Fig. 4 compares the simulation results of the three-layer FSS and the proposed five-layer FSS. As previously mentioned, the three-layer FSS has a passband from 1.944 GHz to 2.562 GHz, and there are two transmission poles in the passband located at 2.064 GHz and 2.418 GHz, respectively. The relative bandwidth of the conventional FSS is 27.4%. It makes sense that the generated broadband is formed by splicing multiple passbands. The vertical incidence simulation results of the proposed five-layer FSS are also shown in Fig. 4b. The passband is from 1.716 GHz to 2.716 GHz. There are three transmission poles in the band located at 1.816 GHz, 2.152 GHz and 2.64 GHz, respectively, and the relative bandwidth is 45.1%. The designed FSS achieves broadband oblique incidence stable transmission of 1.71 – 2.69 GHz with a thickness as small as possible relative to the center frequency wavelength, a relative bandwidth of 44.5%, and a stable angle of 45°, and supports TE and TM dual polarization. The simulation results for the polarization conversion and insertion loss characteristics of the FSS are displayed in Fig. 5. The FSS demonstrates good performance with minimal polarization conversion and insertion loss. The average insertion loss in the band is less than 0.5 dB, which is at a relatively advanced level in similar works.

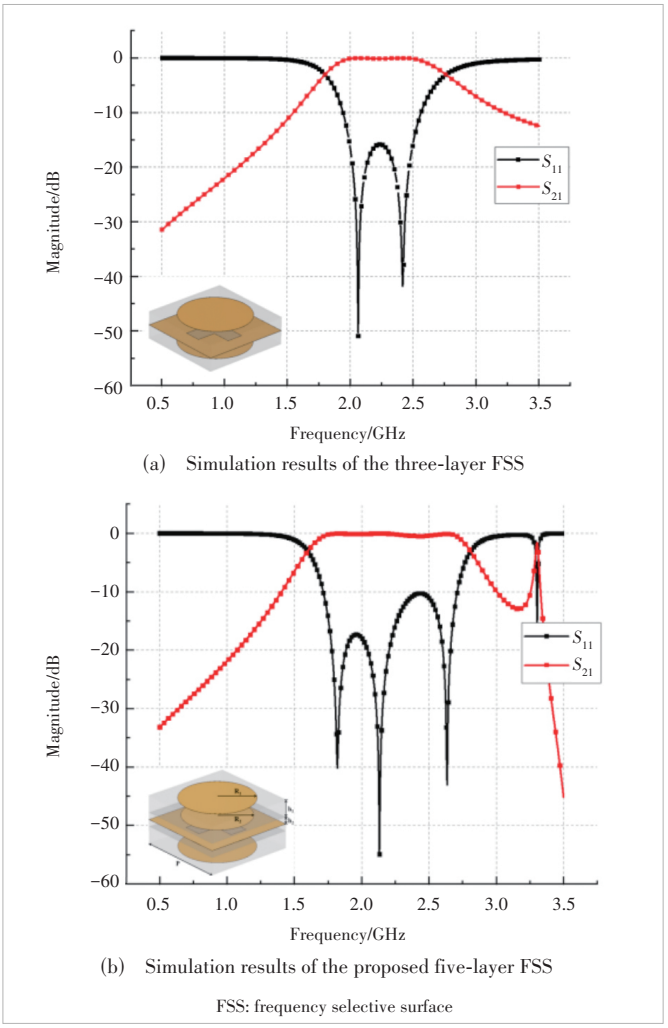


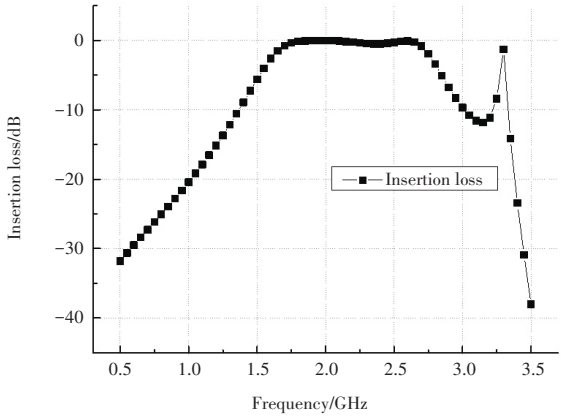
Figure 4. Simulation results of the TE model

To validate the correctness of the designed structure, we extract the equivalent circuit models of the FSS for both the three-layer and five-layer configurations. The results of the

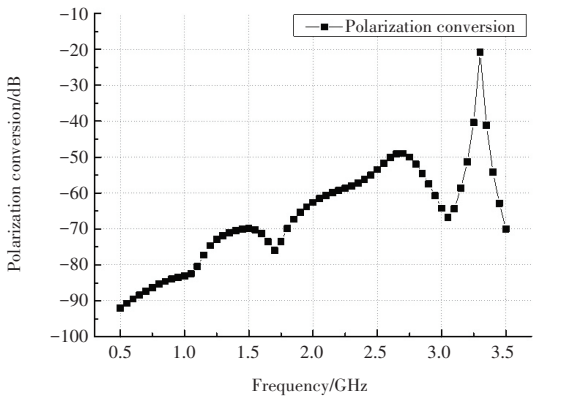
electromagnetic simulations of the equivalent circuits are then compared with those of the circuit simulation. Our proposed structure is reliable and valid, and the results indicate a high level of consistency.

In Fig. 6a, h_1 and h_2 represent two dielectric layers, while Z_{FSS} represents the equivalent circuit of each layer, and they are connected through cascades. In Fig. 6b, L_{m1} represents the equivalent circuit of a circular patch, which is equivalent to an inductance. Considering the coupling relationship with other FSS units, a series capacitor C_{c2} is added; C_{p2} represents the cross gap, equivalent to capacitance; L_{p1} and C_{p1} consist of a series inductor and a parallel capacitor. In addition, it is also necessary to consider the coupling relationship between each layer and add parallel capacitors C_{c1}. The model of the equivalent circuit has been added to Fig. 6.

Different from the three-layer structure, the interlayer coupling effect of the five-layer structure is stronger, so the position of the transmission pole is affected by multiple structural parameters. A more complex model of the five-layer FSS is shown in Fig. 6c.

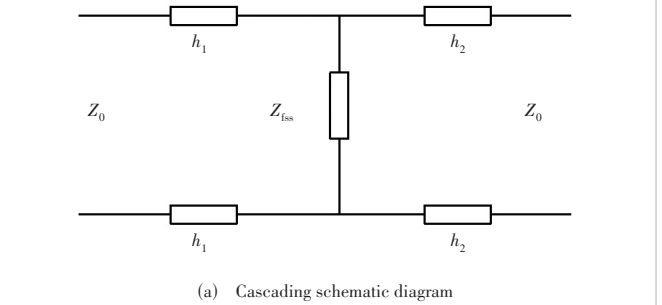


(a) Insertion loss

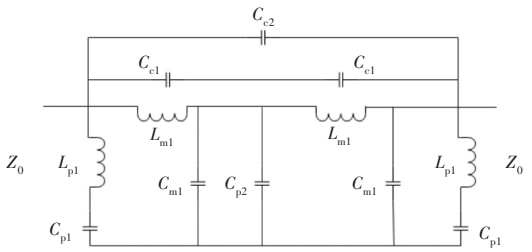


(b) Polarization conversion

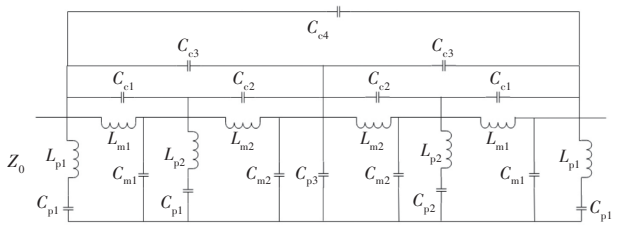
Figure 5. Insertion loss and polarization conversion



(a) Cascading schematic diagram



(b) Three-layer FSS equivalent circuit



(c) Five-layer FSS equivalent circuit

FSS: frequency selective surface

Figure 6. Equivalent-circuit model of FSS

Comparing electromagnetic simulation with equivalent circuits in Fig. 7, we find that equivalent circuits can effectively match the FSS model. Comparing the transmission coefficients of the three-layer FSS and the five-layer FSS, we can see that the five-layer FSS has a wider passband.

The frequency response of the FSS unit cell under oblique incidence is shown in Fig. 8. In Fig. 1, as the incident angle changes, the incident impedance also changes in the opposite direction in TE and TM modes. In the TE mode, as the incident angle increases, the first resonant frequency moves to a lower frequency, and the second and third resonant frequencies move to higher frequency. It can be seen that under the 45° oblique incidence, both modes can provide −10 dB bandwidth from 1.71 GHz to 2.7 GHz. The −0.5 dB transmission coefficient bandwidth of FSS is from 1.71 GHz to 2.68 GHz, and the highest insertion loss in the band is 0.48 dB. Under the 0 – 45° oblique incidence, FSS can support a relative bandwidth of 44.1% and the trans-

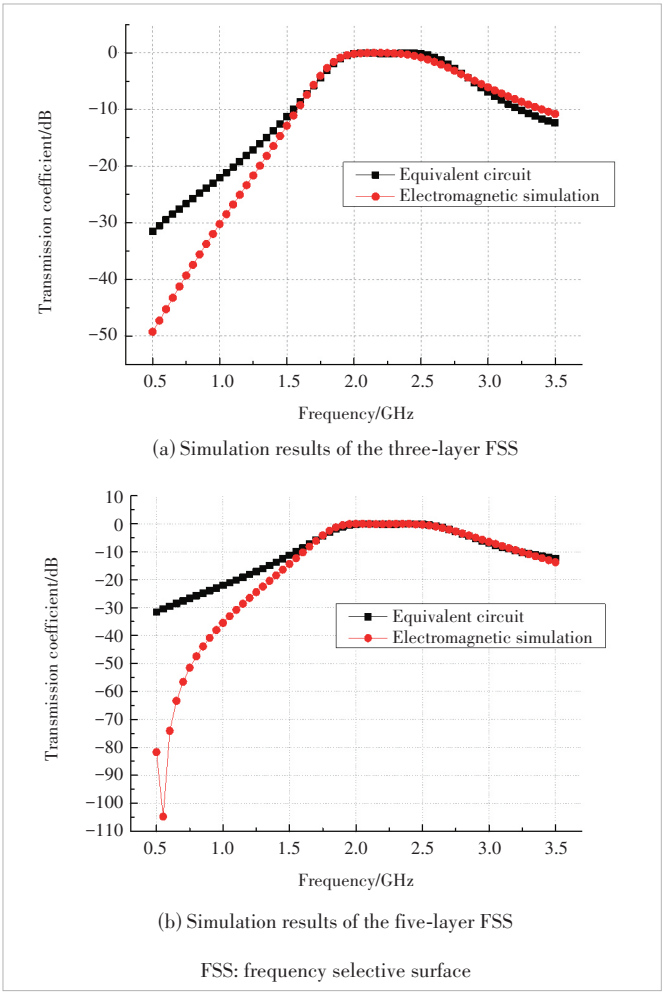


Figure 7. Comparison results of equivalent circuit simulation with electromagnetic simulation

mission coefficient is slightly deteriorated around 1.95 GHz, but still greater than −1 dB. The resonance existing at 3 GHz can be avoided by changing the size of the structure, but doing so will lead to a decrease in performance within the passband. Therefore, we choose to retain the presence of the resonance point at this location.

4 Conclusions

In this paper, a wide passband FSS with angular stability and low insertion loss is proposed. The FSS is designed by a five-layer metal patch stack structure. The upper and lower layers of patches are circular to improve the oblique incidence stability, and the middle layer uses cross slots to control the passband frequency. Simulation results show a wide bandwidth from 1.71 GHz to 2.71 GHz while the insertion loss is less than 0.5 dB. Compared with the three-layer FSS, the designed FSS has a wider passband and better oblique incidence frequency response, which has broad application prospects.

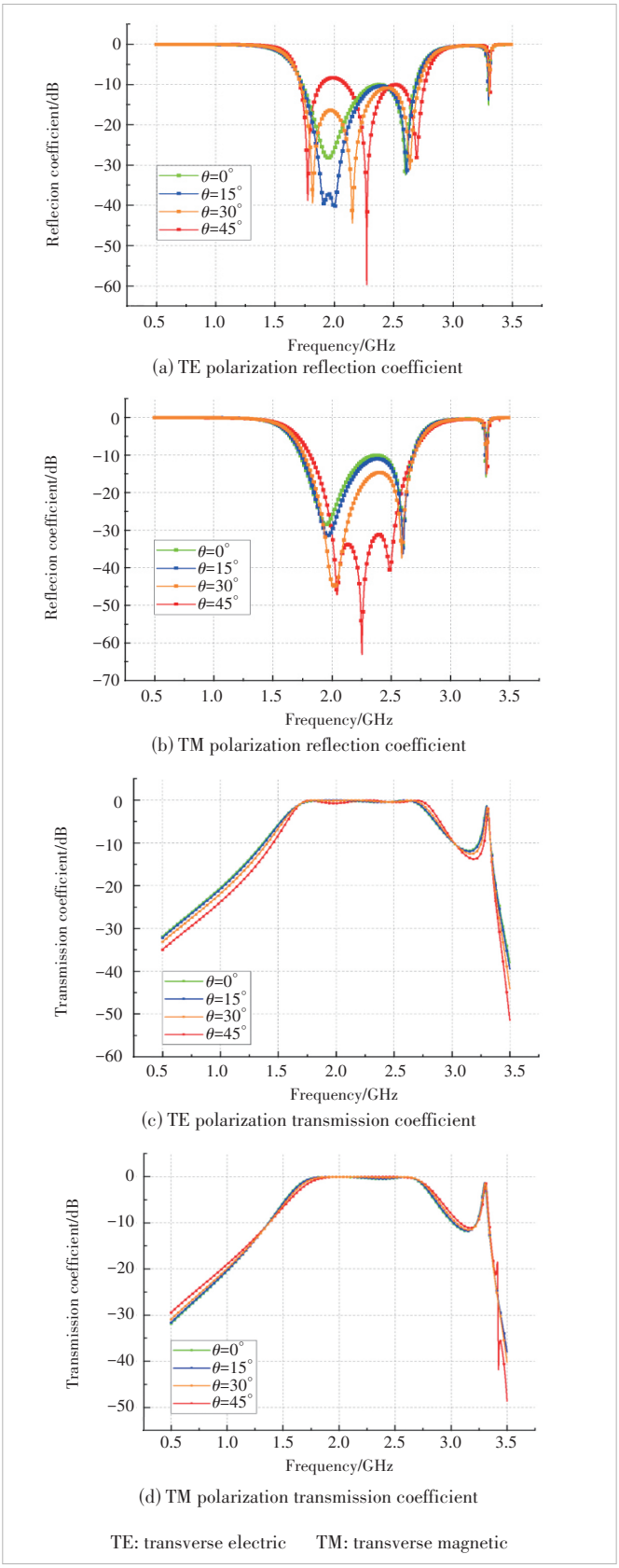


Figure 8. Simulated scattering parameter results of the FSS unit cell under oblique incident angles

References

[1] MITTRA R, CHAN C H, CWIK T. Techniques for analyzing frequency selective surfaces: a review [J]. Proceedings of the IEEE, 1988, 76(12): 1593 - 1615. DOI: 10.1109/5.16352
[2] MUNK B A. Frequency selective surfaces: theory and design [M]. New York, USA: Wiley, 2000. DOI: 10.1002/0471723770
[3] MUNK B A. Finite antenna arrays and FSS [M]. New York, USA: Wiley, 2003. DOI: 10.1002/0471457531
[4] KURRA L, ABEGAONKAR M P, BASU A, et al. FSS properties of a uniplanar EBG and its application in directivity enhancement of a microstrip antenna [J]. IEEE antennas and wireless propagation letters, 2016, 15: 1606 - 1609. DOI: 10.1109/LAWP.2016.2518299
[5] XU R, LI J Y, WEI K, et al. A broadband slot antenna with unidirectional circularly polarized radiation patterns [J]. IEEE antennas and wireless propagation letters, 2016, 16: 317 - 320. DOI: 10.1109/LAWP.2016.2574808
[6] WANG L L, LIU S B, KONG X K, et al. Frequency-selective rasorber with a wide high-transmission passband based on multiple coplanar parallel resonances [J]. IEEE antennas and wireless propagation letters, 2020, 19 (2): 337 - 340. DOI: 10.1109/LAWP.2019.2962223
[7] YANG Z Q, JIANG W, HUANG Q L, et al. A 2.5-D miniaturized frequency-selective rasorber with a wide high-transmission passband [J]. IEEE antennas and wireless propagation letters, 2021, 20(7): 1140 - 1144. DOI: 10.1109/LAWP.2021.3073777
[8] JIN C, LV Q H, WANG J L, et al. Capped dielectric inserted perforated metallic plate bandpass frequency selective surface [J]. IEEE transactions on antennas and propagation, 2017, 65(12): 7129 - 7136. DOI: 10.1109/TAP.2017.2764524
[9] WANG P, JIANG W, HONG T, et al. A 3D wide passband frequency selective surface with sharp roll-off sidebands and angular stability [J]. IEEE antennas and wireless propagation letters, 2022, 21(2): 252 - 256. DOI: 10.1109/LAWP.2021.3126890
[10] WEI P S, CHIU C N, CHOU C C, et al. Miniaturized dual-band FSS suitable for curved surface application [J]. IEEE antennas and wireless propagation letters, 2020, 19(12): 2265 - 2269. DOI: 10.1109/LAWP.2020.3029820
[11] LI T W, FAN Y D, GU Y J, et al. A novel miniaturized multiband strong coupled-FSS structure insensitive to almost all angles and all polarizations [J]. IEEE transactions on antennas and propagation, 2021, 69(12): 8470 - 8478. DOI: 10.1109/TAP.2021.3063351
[12] ZHAO S J, HUANG M, YANG F, et al. A novel wideband FSS radome with loaded inductors [C]//International Applied Computational Electromagnetics Society Symposium (ACES-China). IEEE, 2022: 1 - 2. DOI: 10.1109/ACES-China56081.2022.10065188
[13] DANUOR P, JUNG Y B. A simple reconfigurable FSS structure for antenna beam steering applications [C]//Proceedings of International Conference on Electronics, Information, and Communication (ICEIC). IEEE, 2023: 1 - 3. DOI: 10.1109/ICEIC57457.2023.10049862
[14] LIANG F C, WANG J B, WANG J H, et al. Influence of curvature on the transmission characteristics of cylindrical frequency selective surfaces [J]. Journal of Changchun University of Science and Technology (natural science edition), 2013, 36(Z2): 65 - 66+106

Biographies

TANG Xingyang received his bachelor's degree in electromagnetic field and wireless technology from Harbin Institute of Technology, China in 2022. He is currently pursuing his master's degree in electronic information at Harbin Institute of Technology.

SUI Jia received his BE degree in electronic science and technology from the Harbin Institute of Technology, China in 2021, where he is currently pursuing his ME degree in electronic science and technology. His current research interests are the design of frequency selective surfaces and frequency selective absorber .

FU Jiahui (fjh@hit.edu.cn) received his PhD degree in information and communication engineering from Harbin Institute of Technology, China in 2005. His current research interests are the design of microwave and millimeter wave circuits, antennas, supernormal media, MEMS and EMC. He has published more than 50 relevant academic papers in domestic and foreign journals.

YANG Kaiwen received his BS degree in electronic and information engineering and PhD degree in electromagnetic fields and microwave technology from Xidian University, China in 2016 and 2021, respectively. He is currently an engineer in ZTE Corporation. His current research interests include filtering antennas and base station antennas.

ZHAO Zhipeng received his BS and PhD degrees from Xidian University, China in 2015 and 2020, and then joined ZTE Corporation. He is currently a senior RF system engineer there. His main research directions include the 5G base station, non-terrestrial networks, vehicle radar, etc.

Dual-Polarized 2D Beam-Scanning Antenna Based on Reconfigurable Reflective Elements

LIU Zhipeng[1], LI Kexin[1], CAI Yuanming[1], LIU Feng[2], GUO Jiayin[2]

(1. National Key Laboratory of Antennas and Microwave Technology, Xidian University, Xi'an 710071, China;
2. Department of RHP System, ZTE Corporation, Xi'an 710065, China)

DOI: 10.12142/ZTECOM.202501011

https://kns.cnki.net/kcms/detail/34.1294.TN.20250303.1535.002.html, published online March 4, 2025

Manuscript received: 2023-09-11

Abstract: In this paper, a dual-polarized antenna operating at 3.5 GHz is presented with 2D beam-scanning performance. The steerable beam is realized based on a 2×2 active reflective metasurface. The active metasurface is composed of folded annular rings and cross dipoles embedded with voltage-controlled varactor diodes. By tuning the capacitance values of the varactors, the reflective phase of the metasurface is reconfigured to tilt the main beam. To verify the scanning performance, a prototype is fabricated and measured. At 3.5 GHz, the measured scanning ranges are from −25° to 29° and −27° to 29° in the *XOZ* and *YOZ* planes, respectively.

Keywords: dual-polarized antenna; reflective metasurface; 2D beam scanning; varactor diode

Citation (Format 1): LIU Z P, LI K X, CAI Y M, et al. Dual-polarized 2D beam-scanning antenna based on reconfigurable reflective elements [J]. ZTE Communications, 2025, 23(1): 85 – 89. DOI: 10.12142/ZTECOM.202501011
Citation (Format 2): Z. P. Liu, K. X. Li, Y. M. Cai, et al., "Dual-polarized 2D beam-scanning antenna based on reconfigurable reflective elements," *ZTE Communications*, vol. 23, no. 1, pp. 85 – 89, Mar. 2025. doi: 10.12142/ZTECOM.202501011.

1 Introduction

Beam-scanning antennas have been attracting the interest of investigators for a long time. They have been widely adopted in modern wireless communication systems due to their remarkable ability to balance the requirements of gain and coverage. Compared with omnidirectional antennas, they can obtain a higher gain in a certain desired direction and suppress the interference between neighbor cells. Meanwhile, a steerable beam is capable of satisfying the dynamic distributions of users by time and space. Therefore, numerous ideas about beam-scanning antennas have been proposed.

Among the feasible designs, beam-scanning antennas based on metasurface have attracted great interest from investigators due to the characteristics of light weight, low cost, and ease of fabrication. A mechanical beam-steering antenna using a single-layer passive frequency selective surface (FSS) was presented in Ref. [1], achieving a scanning range of −30° to 30° in the elevation plane via FSS rotation. By applying a Positive-Intrinsic-Negative Diode (PIN)-loaded active metasurface over the square patch antenna[2], the main beam could be switched from −30° to 30° with a faster response. In Ref. [3], a 1D beam-switching antenna was proposed based on a PIN-loaded reflector. The maximum tilting angle is 30° by controlling the states of diodes. In Ref. [4], a 2D beam-switching antenna was presented at 5.5 GHz based on a 6×6 reconfigurable partially reflective surface (PRS). By controlling the states of PIN diodes in different sections of PRS, a ±47° beam switch could be achieved in the azimuth plane. Based on the cylindrical metasurface surrounding an active dipole, a 360° beam horizontal sweeping and a discrete elevation switching between −22° and 22° were achieved in Ref. [5]. To achieve continuous beam scanning, a novel phased array was proposed in Ref. [6] for 5G millimeter-wave wireless communications. Different from the traditional phased arrays that employ phase shifters to electronically control the beam direction, the ±60° scanning beam was realized based on a 256-element active electromagnetic (EM) surface fed by a horn.

Likewise, a 196-element reflective metasurface in Ref. [7] was used to change the beam direction of an antenna, which enabled continuous beam scanning within ±20° by controlling PIN diodes loaded on phase delay lines. In Ref. [8], a varactor-controlled reflective metasurface fed by a monopole was designed to steer the beam within ± 50° in elevation directions

This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. HC-CN-20220719005.

in a single polarization. For dual-polarized use, Ref. [9] proposed a beam-scanning cross dipole antenna loaded with an active reflective metasurface with fewer component counts, achieving ±20° beaming scanning via varactor reconfiguration.

In this paper, a beam-scanning antenna is presented based on a 2×2 active reflective metasurface. For dual-polarization use, the surface is excited by a printed cross dipole. By tuning the voltages applied across the varactors, the reflective phases of the elements can be reconfigured thus deflecting the beam direction in the *XOZ* and *YOZ* planes.

2 Design of Beam-Scanning Antenna

2.1 Reconfigurable Reflective Element

The proposed reconfigurable reflective element is illustrated in Fig. 1. The element consists of two Rogers RO4003 dielectric substrates (ε_r=3.55, tanδ = 0.002 7) with a 6 mm thick air spacing between them. A cross and a cross-shaped ring are separately printed on the top and bottom sides of the upper dielectric substrate with a thickness of 0.508 mm. Four varactors are embedded in the cross to adjust the equivalent electrical length of the cross strips. The capacitance values of the four varactors are equal to C_1. As a result, the resonance frequency of the cross is tunable with the varactor value C_1 changing. Four inductors and metal vias are used to connect the cross strips to the cross-shaped folded annular ring. Compared to a square ring, the cross-shaped design lengthens the path of the surface current without expanding the transversal dimensions. The ground plane is designed on the top side of the substrate with a thickness of 0.813 mm. The element features a symmetric structure for polarization-insensitive frequency response.

The proposed reflective element is modeled in the High-Frequency Structure Simulator (HFSS) and simulated with periodic boundaries. According to the parameters of Skyworks SMV1430 varactor diodes, the varactors are modeled as a tunable capacitor in series with a 3 Ω resistor and a 0.45 nH inductor. Fig. 2 shows the simulated results of the reflective amplitude and phase response. A smooth reflective phase response is observed from 2.5 GHz to 4.5 GHz. When the capacitance value C_1 changes from 0.31 pF to 1.1 pF, the phase curve moves to a lower band thus changing the reflective phase in the operated band. In other words, there is a phase shifting between the reflected and incident waves due to the phase compensation by the element. As Fig. 2 shows, the phase shifting at 3.5 GHz is about 214° with a reflective amplitude no lower than −2 dB. The cross-polarization component of the reflected wave remains below −27 dB.

2.2 Beam Steering Antenna Design

To operate the varactors, the DC bias circuit is designed as shown in Fig. 3. The varactor anodes loaded on each element are connected via the cross strip. A central grounding metal

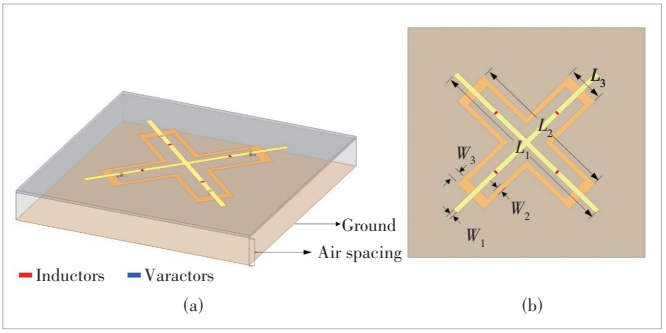


Figure 1. Geometry of the proposed reconfigurable reflective element: (a) 3D view; (b) top view (L_1=38.5 mm, L_2=30 mm, L_3=7 mm, W_1=1 mm, W_2=1 mm, and W_3=2 mm)

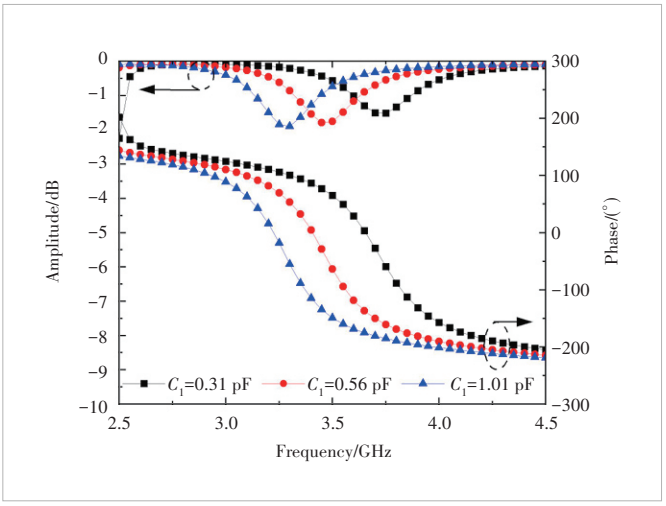


Figure 2. Reflective amplitude and phase response versus frequencies at different capacitance values of each varactor

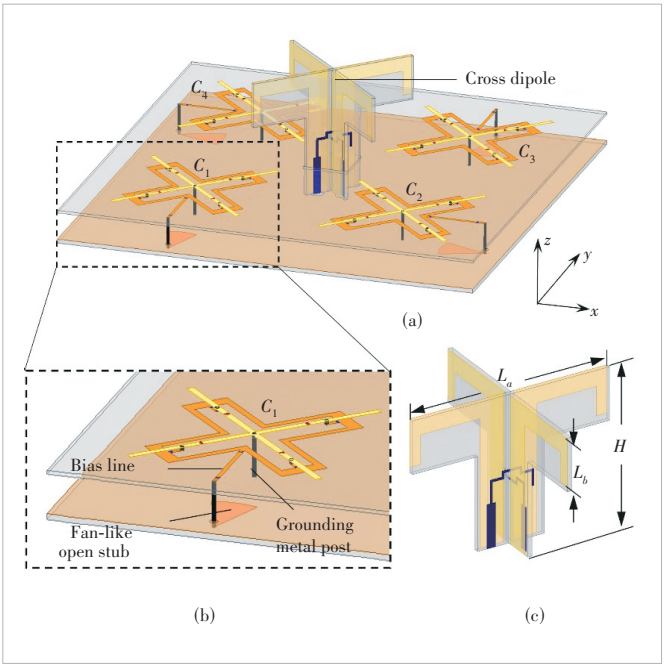


Figure 3. Geometry of the proposed antenna: (a) 3D view; (b) structure of bias circuit for varactors; (c) printed cross dipole

post with a diameter of 0.6 mm sets the anode potential to zero. The cathodes of the varactor diodes are connected by the cross-shaped ring to share a common biasing point. Another metal post is designed through the ground plane at a distance of 17.9 mm away from the grounding post. A hole is etched on the ground plane to isolate metal posts from the ground plane. At the bottom end of the post, a fan-like open stub acts as a low-pass filter for isolating the RF signal from the DC bias signal. Together with the post, the bias line stretching from the cross-shaped ring is designed to offer positive electrical potential to the cathodes of the diodes. As a result, the four shunt diodes are reverse biased with the same capacitance value.

To achieve a 2D scanning beam, the proposed element is arranged along the x and y directions to form a 2×2 reflective surface. A printed cross dipole operating at 3.5 GHz radiates ±45° polarization electromagnetic waves and feeds the surface. The four elements are set surrounding the cross dipole and share a common ground. Fig. 3 shows that the bias structures of the elements are designed in central symmetry. The overall size of the proposed antenna is about $1.05\lambda_0 \times 1.05\lambda_0 \times 0.3\lambda_0$, where λ_0 is the 3.5 GHz free-space wavelength. The capacitance values of different elements are C_1, C_2, C_3, and C_4 respectively (Fig. 3). Therefore, four independent DC bias voltages are needed. When the bias voltage applied across the varactors of each element varies, the re-radiating phase of the reflective element is changed. According to the in-phase superposition principle, the wavefront will incline thus letting the beam squint. In summary, beam scanning can be achieved by tuning the bias voltages of the reflective surface.

3 Simulation and Measurement Results

The proposed antenna is simulated, fabricated, and measured to further verify its scanning performance, as shown in Fig. 4. Nylon support components are used to obtain a 6 mm thick air spacing between the reflective surface and the common ground. The desired values of C_1, C_2, C_3, and C_4 are controlled by corresponding voltages U_1, U_2, U_3, and U_4. The proposed antenna is simulated and measured at five states as listed in Table 1.

Fig. 5 shows the simulated and measured S parameters for different states. Within the operated band from 3.4 GHz to 3.6 GHz, the measured S_{11} values are less than −11.0 dB, while the simulated S_{11} is less than −11.2 dB.

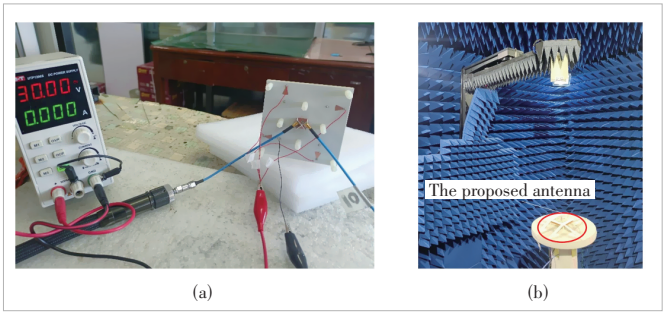


Figure 4. Fabricated antenna prototype under measurements: (a) S parameter measurement setup and (b) radiation pattern measurement setup

Table 1. Beam states of the proposed antenna and the setup of varactors

Beam State	Varactor Capacitance/pF				Varactor Biasing Voltage/V			
	C_1	C_2	C_3	C_4	U_1	U_2	U_3	U_4
I	0.31	0.31	0.31	0.31	30	30	30	30
II	0.31	1.10	1.10	0.31	30	0	0	30
III	1.10	0.31	0.31	1.10	0	30	30	0
IV	0.31	0.31	1.10	1.10	30	30	0	0
V	1.10	1.10	0.31	0.31	0	0	30	30

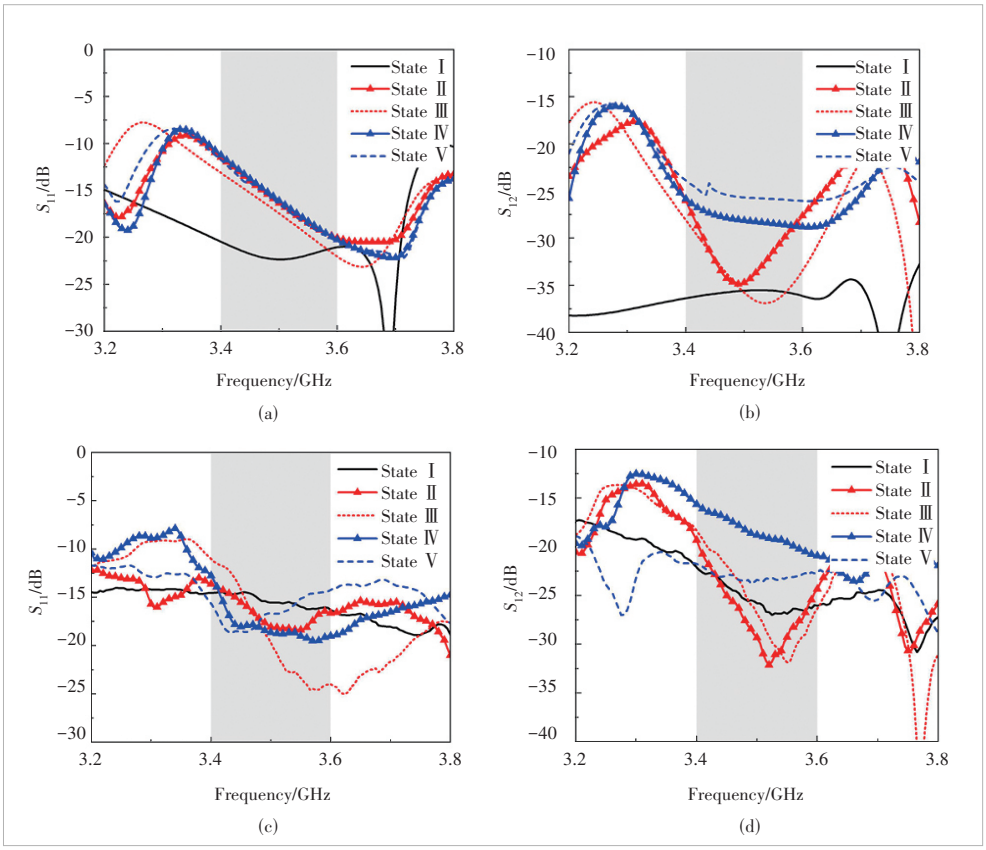


Figure 5. S parameters at different beam states: (a) simulated S_{11}; (b) simulated S_{12}; (c) measured S_{11}; (d) measured S_{12}

The measured S_{12} is less than −15.6 dB, while the simulated S_{12} is less than −23.8 dB.

The radiation patterns for various steering states were measured using a near-field antenna measurement system in an anechoic chamber. Fig. 6 shows the simulated and measured radiation patterns for 45° polarization in both the *XOZ* and *YOZ* planes when Port 1 is excited. By reconfiguring the bias voltages, the main beam can be deflected as anticipated.

When all bias voltages are set to 30 V, the main beam points directly upwards without beam steering, corresponding to State I. By adjusting U_2 and U_3 to 0 V, the measured steering angle at 3.5 GHz is −25°, which corresponds to State II. When bias voltages U_1 and U_4 are tuned to 0 V, while U_2 and U_3 remain at 30 V, the antenna operates in State III with a steering angle of 29° in the *XOZ* plane. Similarly, when the bias voltages are configured for States IV and V, the maximum scanning range in the *YOZ* plane spans from −27° to 29°. The measured results align well with the simulated ones, indicating the accuracy and reliability of the proposed antenna design.

4 Conclusions

In this paper, a continuous beam steering antenna based on the 2×2 active reflective metasurface is modeled and fabricated. By changing the voltages applied on the active reflective metasurface, the proposed antenna can steer the beam in both the *XOZ* and *YOZ* planes in the frequency range of 3.4 – 3.6 GHz. The measured reflection coefficient is less than −10 dB and the port isolation is greater than 15 dB. The measured scanning ranges are −25° to 29° and −27° to 29° in the *XOZ* and *YOZ* planes, respectively. The antenna is a good candidate for application to beam reconfigurable communication systems.

References

[1] DAS P, MANDAL K, LALBAKHSH A. Beam-steering of microstrip antenna using single-layer FSS based phase-shifting surface [EB/OL]. (2021-12-14)

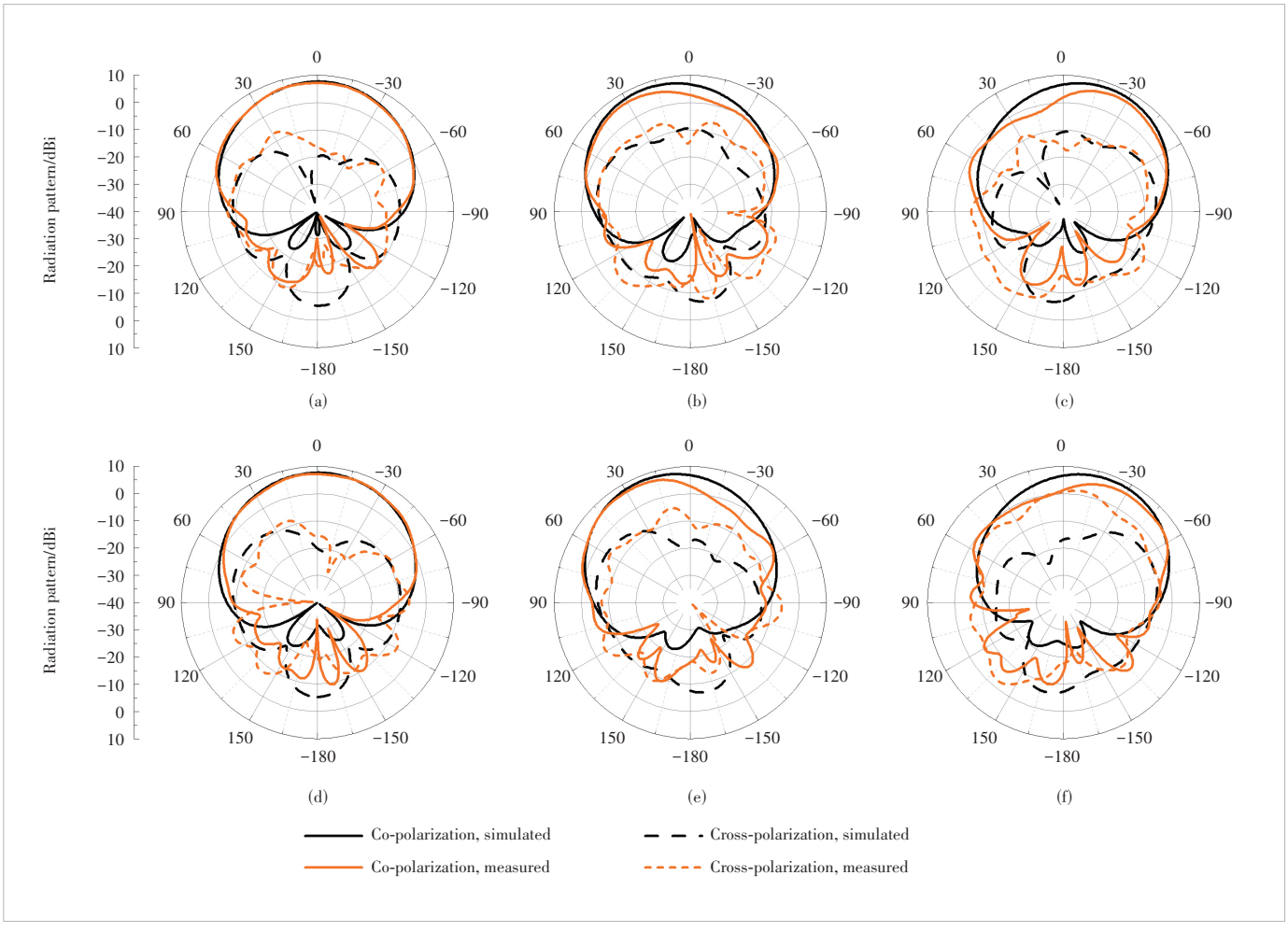


Figure 6. Radiation patterns of different beam states: (a) State Ⅰ in the *XOZ* plane; (b) State Ⅱ in the *XOZ* plane; (c) State Ⅲ in the *XOZ* plane; (d) State Ⅰ in the *YOZ* plane; (e) State Ⅳ in the *YOZ* plane; (f) State Ⅴ in the *YOZ* plane

[2023-05-01].https://onlinelibrary.wiley.com/doi/10.1002/mmce.23033
[2] SURYAPAGA V, KHAIRNAR V V. Pattern reconfigurable antenna using programmable metasurface [C]//Proc. 3rd International Conference on Artificial Intelligence and Signal Processing (AISP). IEEE, 2023: 1 – 5. DOI: 10.1109/AISP57993.2023.10134935
[3] MAJUMDER B, MUKHERJEE J, KRISHNAMOORTHY K, et al. A novel beam steering dipole antenna using phase varying metasurface as reflector [C]//Proc. IEEE International Conference on Antenna Innovations & Modern Technologies for Ground, Aircraft and Satellite Applications (iAIM). IEEE, 2017: 1 – 4. DOI: 10.1109/IAIM.2017.8402618
[4] NADEEM M, SHOAIB N, RAZA A, et al. 2-dimensional (2D) beam steering-antenna using active PRS for 5G applications [J]. Micromachines, 2022, 14(1): 110. DOI: 10.3390/mi14010110
[5] SURIER A, HAKEM N, KANDIL N. 3D beam steering cylindrical antenna [C]//Proc. IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting (AP-S/URSI). IEEE, 2022: 1588 – 1589. DOI: 10.1109/AP-S/USNC-URSI47032.2022.9886357
[6] LI Y Z, REN Y L, YANG F, et al. A novel 28 GHz phased array antenna for 5G mobile communications [J]. ZTE communications, 2020, 18(3): 20 – 25. DOI: 10.12142/ZTECOM.202003004
[7] WANG Z L, GE Y H, PU J X, et al. 1 bit electronically reconfigurable folded reflectarray antenna based on p-i-n diodes for wide-angle beam-scanning applications [J]. IEEE transactions on antennas and propagation, 2020, 68(9): 6806 – 6810. DOI: 10.1109/TAP.2020.2975265
[8] NAM I J, LEE S, KIM D. Miniaturized beam reconfigurable reflectarray antenna with wide 3-D beam coverage [J]. IEEE transactions on antennas and propagation, 2022, 70(4): 2613 – 2622. DOI: 10.1109/TAP.2021.3083732
[9] JIA Y T, JIANG G S, LIU Y, et al. Beam scanning for dual-polarized antenna with active reflection metasurface [J]. IEEE antennas and wireless propagation letters, 2022, 21(9): 1722 – 1726. DOI: 10.1109/LAWP.2022.3176427

Biographies

LIU Zhipeng received his BS degree from Xidian University, China in 2021. He is currently pursuing his master's degree in electromagnetic wave and microwave technology at Xidian University. His current research interests include reconfigurable metasurface antennas and base station antennas.

LI Kexin received his BS degree from Shandong University of Science and Technology, China in 2021. He is currently pursuing his master's degree in new-generation electronic information technology at Xidian University, China. His current research interests include phased array design and reconfigurable antenna design.

CAI Yuanming (ymcai@xidian.edu.cn) received his BS degree in electronic information engineering and PhD degree in electromagnetic wave and microwave technology from Xidian University, China in 2011 and 2016, respectively. He is currently an associate professor with the National Key Laboratory of Radar Detection and Sensing, the School of Electronic Engineering, Xidian University. His current research interests include multiband and wideband antennas, circularly polarized antennas, and reconfigurable antennas.

LIU Feng received his BS and PhD degrees from Xidian University, China in 2016 and 2021, respectively. He has been a senior RF system engineer in the RCH System Design Department of ZTE Corporation since 2021.

GUO Jiayin received her BS degree in electronic information engineering and PhD degree in electronic science and technology from Xidian University, China in 2016 and 2022, respectively. She has been a senior RF engineer in the RCH System Design Department of ZTE Corporation since 2022.

VFabric: A Digital Twin Emulator for Core Switching Equipment

WANG Qianglin[1], ZHANG Xiaoning[1], YANG Yi[1], FAN Chenyu[1], YUE Yangyang[2], WU Wei[2], DUAN Wei[2]
(1. School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 610000, China; 2. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202501012
https://kns.cnki.net/kcms/detail/34.1294.TN.20250319.0908.002.html, published online March 19, 2025
Manuscript received: 2023-10-24

Abstract: The proliferation of heterogeneous networks, such as the Internet of Things (IoT), unmanned aerial vehicle (UAV) networks, and edge networks, has increased the complexity of network operation and administration, driving the emergence of digital twin networks (DTNs) that create digital-physical network mappings. While DTNs enable performance analysis through emulation testbeds, current research focuses on network-level systems, neglecting equipment-level emulation of critical components like core switches and routers. To address this issue, we propose vFabric (short for virtual switch), a digital twin emulator for high-capacity core switching equipment. This solution implements virtual switching and network processor (NP) chip models through specialized processes, deployable on single or distributed servers via socket communication. The vFabric emulator can realize the accurate emulation for the core switching equipment with 720 ports and 100 Gbit/s per port on the largest scale. To our knowledge, this represents the first digital twin emulation framework specifically designed for large-capacity core switching equipment in communication networks.

Keywords: digital twin network; core switch/router; sockets; network emulation

Citation (Format 1): WANG Q L, ZHANG X N, YANG Y, et al. VFabric: a digital twin emulator for core switching equipment [J]. ZTE Communications, 2025, 23(1): 90 - 100. DOI: 10.12142/ZTECOM.202501012
Citation (Format 2): Q. L. Wang, X. N. Zhang, Y. Yang, et al., "VFabric: a digital twin emulator for core switching equipment," *ZTE Communications*, vol. 23, no. 1, pp. 90 - 100, Mar. 2025. doi: 10.12142/ZTECOM.202501012.

1 Introduction

In recent years, with the fast development of information and communication technology (ICT), such as big data, cloud computing, and artificial intelligence (AI), there has been a rise in emerging heterogeneous communication networks, such as the Internet of Things (IoT)[1], unmanned aerial vehicle (UAV) networks[2], and edge networks[3]. These different network forms have complex operation and administration requirements. For example, UAV networks should effectively adapt to dynamic topology due to UAV flight and satisfy the quality of service (QoS) requirements of various types of traffic[4]. Meanwhile, the number of nodes or mobile devices connected to communication networks is increasing explosively, which brings scalability and flexibility challenges. In general, current communication networks have become increasingly complex and difficult to operate and manage.

To solve this problem, the digital twin network (DTN) technology has been introduced to facilitate the effective management of communication networks[5]. The concept of digital twin (DT) has been developed in many industries for decades. Today, the DT technology has been widely applied in a large variety of domains, including smart manufacturing Industry 4.0[6], aviation[7], healthcare[8], communication networks[9], and smart grid systems[10]. The basic idea of DT is a digital representation or virtual model of a single physical object. It is a system that focuses on producing a virtual model of a physical entity with high fidelity. Such a system needs to be intelligent and persistently evolving[11]. A DT system generally contains three main modules: a physical object in physical space, a virtual object in virtual space, and the data connection between the two spaces. Leveraging the DTN technology, a high-fidelity emulation system is established for efficiently controlling and managing dynamic and complex communication networks. The DTN emulation system allows network operators to analyze and forecast network performance, develop network solutions, precisely pinpoint network failures, and upgrade networks to accommodate the demands of a growing user base and the integration of new technologies[12].

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant Nos. 62171085, 62272428, 62001087, U20A20156, and 61871097 and the ZTE Industry-University-Institute Cooperation Funds under Grant No. HC-CN-20220722010.

According to the definition and four core elements of the digital twin network[13], it can be designed as a "three-layer three-domain dual-closed-loop" architecture (Fig. 1). The three layers include the physical network layer, the twin network layer, and the network application layer, while the three domains correspond to the data domain, model domain, and management domain of the twin network layer. These domains are respectively implemented by the data-sharing repository, service mapping model, and network twin management subsystems. Meanwhile, the "dual-closed-loop" refers to the inner closed-loop optimization based on the service mapping model and the outer closed-loop control, feedback, and optimization based on the three-layer architecture. The physical network layer is a component of the digital twin network, where various network elements exchange network data and control information with the network twin body through the twin southbound interface. The twin network layer is the hallmark of the digital twin network system, containing three critical subsystems: the data sharing repository, service mapping model, and network twin management subsystems. The network application layer controls the digital twin network. Network applications input requirements to the twin network layer through the twin northbound interface and deploy services in the twin network layer via model instances.

In recent years, some research works have been conducted on applying digital twin technology in the field of communication networks (i.e., DTN). DTN is a key enabler for efficient management in communication networks. In particular, the virtual models of DTN can reflect the dynamic characteristics of physical communication networks (e.g., dynamic network topology, growing traffic flows, or devices)[14]. Thus, the network administrators can effectively manage the network considering the network dynamics. For example, the network administrators can easily perform network planning and complete traffic engineering with the help of DTN technology. Therefore, the DTN can accurately forecast the future network state and provide optimal solutions. On the other hand, current DTN research works always study the entire network system (such as 6G networks, vehicular networks, and the IoT)[15] and do not investigate the virtual model of communication network equipment. In fact, how to build an accurate DT virtual model for communication network equipment, especially high-capacity core network switches or routers, is an important issue to be studied. Since the core switch or router has a large-capacity (i.e., a large number of ports, and each port with at least a rate of 100 Gbit/s), achieving high-fidelity emulation for large-capacity communication equipment is a great technical challenge.

To address this issue, we develop a digital twin emulator for large-capacity core switching equipment, called vFabric

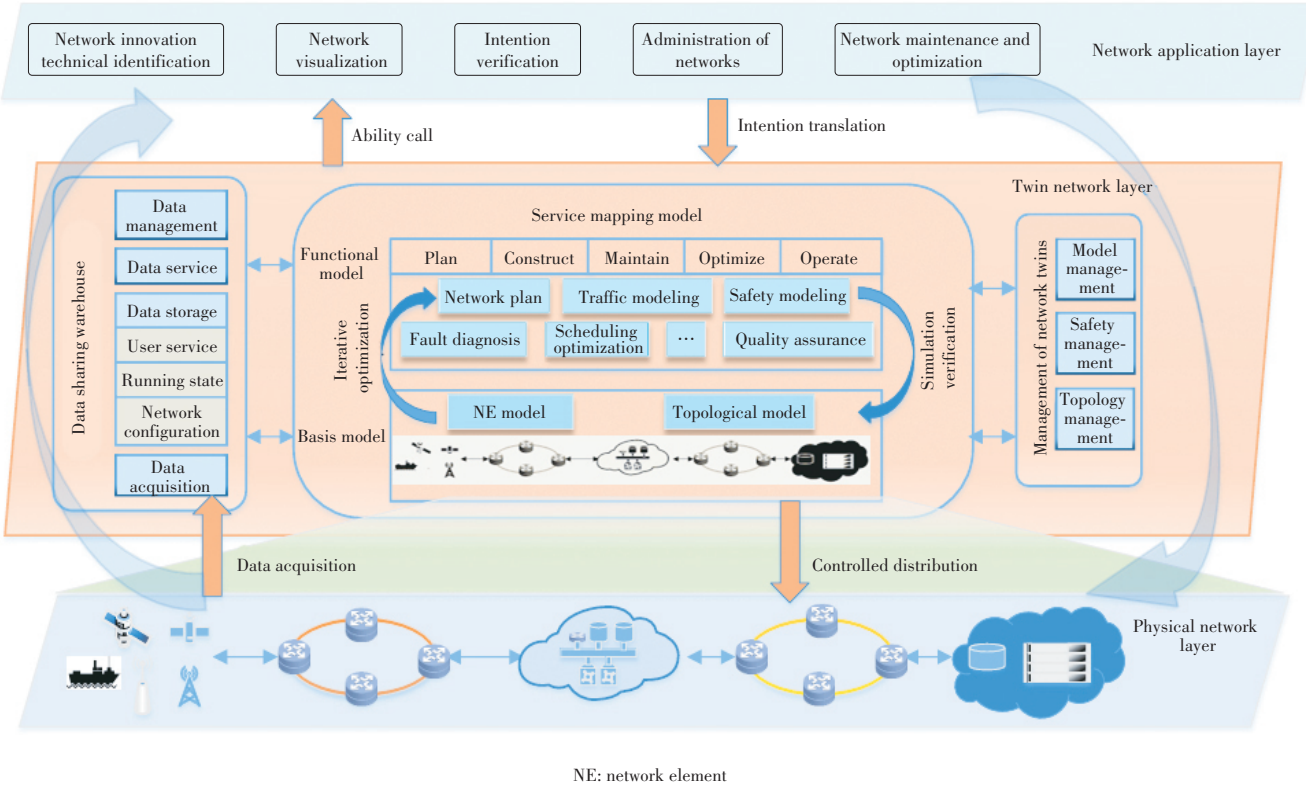


Figure 1. Architecture of digital twin network

(short for "virtual switch"), which enables performance troubleshooting and optimization. When a specific system module fails in practice, the corresponding module in vFabric can be analyzed to identify possible faults, quickly locate the problem, and provide fault diagnosis. At the same time, when a system performance bottleneck occurs, various design schemes can be verified and tested in vFabric, and their performance, reliability, and efficiency can be evaluated to ensure the selection of the optimal design. We use the vFabric emulator to simulate the Clos network inside the core switching equipment and build virtual models for switching chips and network processor (NP) chips, which are key components of the core switch/router. For small-scale core switching equipment, the vFabric emulator is implemented on a single physical server. For large-scale core switching equipment, the vFabric emulator is distributed across multiple physical servers interconnected via sockets. Our vFabric emulator achieves accurate emulation of core switching equipment with up to 720 ports and 100 Gbit/s per port on the largest scale. To the best of our knowledge, this is the first study on DT emulation for large-capacity core switching equipment in the field of communication networks. Our work addresses the gap in digital twin applications for communication devices and pioneers the application of DT to communication devices.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the system model and technical challenges, while Section 4 details the design of vFabric. Section 5 presents the implementation of vFabric and Section 6 provides the emulation results. Section 7 concludes the paper.

2 Related Work

Digital twins have been applied in various domains. Here, we only review important areas such as the Internet of Vehicles (IoV), edge networks, and 6G.

1) IoV: The digital twin-assisted decision-making framework for the IoV leverages the integration of communication, sensing, and computing to enhance vehicle collaboration. FU et al. [16] introduced a digital twin technology concept that maps vehicles and roadside infrastructure from physical space to cyberspace to form simulated and reconstructed virtual entities. Ref. [16] also discussed a multi-agent system (MAS) approach to modeling connected autonomous driving, using artificial intelligence algorithms such as deep reinforcement learning (DRL) to enable decision-making. It highlighted the limitations of traditional multi-agent deep reinforcement learning (MADRL) methods, where agents are not connected with each other, and emphasized the importance of enabling agent-to-agent communications. QIN et al. [17] investigated pricing strategies and resource management between vehicles and mobile edge network (MEC) servers when combining digital twins and MEC in the IoV. Ref. [17] also established a dynamic digital twin for the air-assisted IoV to capture time-varying resource supply and demand, enabling unified resource scheduling and allocation.

2) Edge networks: PILLAI et al.[18] implemented a DT system in vehicular networks to enhance edge computing capabilities. The DT collected data from roadside units (RSUs) and optimized task offloading and resource allocation for efficient system load management. DAI et al.[19] integrated digital twin technology into vehicular edge computing networks to improve network management and offloading efficiency. The proposed adaptive digital twin-enabled network utilized virtual representations of the physical network, and a deep reinforcement learning-based offloading scheme was designed to minimize latency[19]. DAI et al.[20] explored a DTN-assisted MEC system, aiming to maximize the number of service requests served by MECs or minimize the load on the cloud. GUO et al.[21] focused on utilizing digital twin technology to enhance the management efficiency of physical entities in edge computing networks. The proposed mechanism included a time-frequency correlation-based activity estimation model and a chaotic particle swarm optimization algorithm for network sensing edge deployment.

3) 6G: NJOKU et al.[22] explored the potential application of digital twin technology in 6G communication systems. They emphasized the need for innovative architectures and enabling technologies to meet the demanding requirements of 6G systems. TAO et al.[23] proposed a software-defined DTN architecture with virtualization for adaptive 6G service response. They also introduced a deep reinforcement learning-based resource orchestration algorithm to optimize service quality. LU et al.[24] focused on integrating digital twin technology with edge networks to address the challenges in building 6G networks with ubiquitous connectivity, low latency, and enhanced edge intelligence.

4) Data communication networks: WEI et al.[25] discussed data-driven routing, a typical network function under the DTN framework, and demonstrated the potential of DTNs to solve traditional network problems. In SDN-based networks, RAJ et al.[26] proposed a data representation-based DTN architecture that integrates knowledge graphs (KGs) for data modeling and storage. ONO[27] et al. presented a scheme called Area-Controlled Mobile Ad-Hoc Networking (AMoND). The digital twin used in AMoND focuses on managing node location information and does not need to fully replicate real-world environments on a computer.

While the related works presented above focus on building DTN models for entire network systems (such as 6G and edge networks) without considering the virtual modeling of communication network equipment, our work specifically addresses high-fidelity emulation of large-capacity core switching equipment. To our knowledge, this is the first study on developing a DT model for core switches/routers.

3 System Model and Technical Challenges

As an important transmission and forwarding device, large-capacity core switching equipment (e.g., the core routers Huawei NetEngine 8000[28], Cicso 8000[29], and T8000[30], and the core switch ZXR10 9900/9900-S[31]) has been widely deployed in various scenarios including backbone networks, campus networks, data center networks, etc. To satisfy the requirement of high bandwidth and intelligent management, large-capacity core switching equipment has the following characteristics: high reliability, scalability, and performance. These features enable enhanced network bandwidth, elimination of bottlenecks, congestion mitigation, and support for diverse traffic interfaces[32]. Therefore, the performance of large-capacity core switching equipment directly impacts the stability and QoS of the whole communication network.

3.1 Logic Model of Large-Capacity Core Switching Equipment

Generally, large-capacity core switching equipment comprises four fundamental modules (Fig. 2): the input module, output module, switching fabric, and control module. The switching fabric is composed of multiple basic switching units[33]. The first three modules constitute the data plane of switching equipment, while the last module belongs to the control plane. The data plane is responsible for packet forwarding, while the control plane is responsible for generating the data path in the switching fabric. During packet forwarding in large-capacity core switching equipment, the control module first calculates the routing path of the incoming packets, i.e., generating the Routing Information Base (RIB). Subsequently, the Forwarding Information Base (FIB) is generated from the RIB and installed in basic switching units of the switching fabric module. This ensures accurate transmission of incoming packets from the input port to the output port through the switching fabric. Inside large-capacity core switching equipment, the switching fabric and its corresponding scheduling algorithm play an important role in switching performance (throughput and delay). At present, the frequently used switching fabrics are single-stage crossbar[34] and multistage Clos networks[35]. Since the number of ports and the read/writing rate of shared memory limit the performance of single-stage crossbar, the multistage Clos network is more broadly used than the single-stage crossbar. When the number of ports increases, compared with the single-stage crossbar, the multistage Clos network can effectively reduce the number of crossover nodes by an order of magnitude. In a multistage Clos network, two basic switching units are connected by only one link, but multiple paths exist between the arbitrary input port and output port[35]. Therefore, the multistage Clos network can support multipath transmission and achieve load balancing for traffic. On the other hand, in the multistage Clos network, the basic switching units of each stage have the same scale (the same number of input and output ports), which means it has good scalability. Thus, small-scale basic switching units can be used to construct a large-scale/capacity Clos network. In summary, the multistage Clos network has the following advantages: modularity, non-blocking with multiple paths, and good scalability. These features make it a preferred choice for commercial off-the-shelf (COTS) core routers or switches[36].

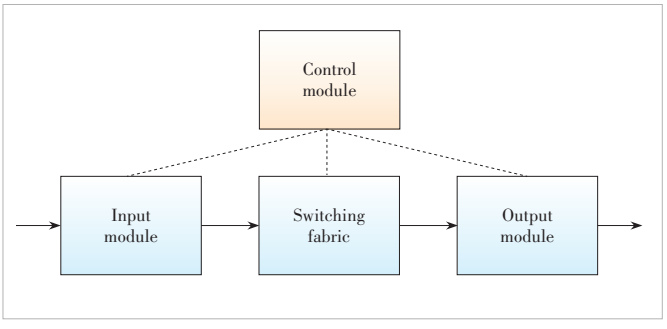


Figure 2. Logic model of large-capacity switching equipment

The three-stage Clos network consists of an input stage, an intermediate stage, and an output stage. The input stage comprises k $n\times m$ crossbars, where n denotes the number of input ports of each crossbar and m represents the number of output ports of each crossbar. The intermediate stage consists of m $k\times k$ crossbars, where k denotes the number of input and output ports of each crossbar. The output stage is composed of k $m\times n$ crossbars, where m denotes the number of input ports of each crossbar and n represents the number of output ports of each crossbar. Therefore, the three-stage Clos network can be denoted as $C(n, m, k)$.

3.2 Hardware Model of Large-Capacity Core Switching Equipment

Based on the logic model of large-capacity core switching equipment, we further illustrate its hardware model. The hardware model of core switching equipment is composed of a main control board, a service board, and a switching board. The main control board corresponds to the control module in the logic model, which generates the routing path inside the switching equipment by the control software, such as the Open Shortest Path First (OSPF) protocol, running on the Linux operating system, and sends the FIB to the switching chips on the data plane. The main control board communicates with the service board and the switching board with socket network communication over Ethernet. The service board comprises the NP chip, the interface chip, and the switching chip. The main work of the NP chip is to perform traffic flow scheduling and QoS management. It is important that each port in the service board is bidirectional, serving as both an input and an output port. The switching board consists solely of the switching chip, which is responsible for high-speed switching. Key components of the switching chip include the Parser (for parsing packet headers), the Forwarding Table (for recording

packet forwarding paths), and the Buffer (for storing packets during switching), among others.

3.3 Technical Challenges for DT Model

Since large-capacity core switching equipment is a key enabler for various types of communication networks, e.g., backbone networks, campus networks, and data center networks, it is necessary to develop a DT model for such equipment[37]. This allows network researchers and operators to test and verify the new technology in the DT model and obtain accurate emulation results. Undoubtedly, the DT model for core switching equipment can help design network optimization algorithms, analyze network performance, and forecast network status under new traffic patterns.

The main technical challenge in developing a DT model for high-capacity core switching equipment is accurately emulating the high bandwidth and large traffic environment of core routers/switches. Since the core switching equipment comprises a large number of service boards and switching boards (normally having 20 ports, each supporting 100 Gbit/s at least), its DT model needs to emulate large volumes of traffic, which brings a great challenge for the existing simulation tools (e.g., OPNET, OMNeT++, and NS-3). Another important issue is the scalability of the core switching equipment. By combining varying numbers of service boards and switching boards, larger-scale core switching equipment can be created, which enhances the emulation difficulty of the DT model.

4 Design of VFabric

In this paper, to address the aforementioned challenges, we propose vFabric, a digital twin emulator for large-capacity core switching equipment. The name "vFabric" is derived from "virtual switch", reflecting its purpose. Our vFabric is a distributed architecture with high scalability and accuracy. It not only simulates the details of the core switch, but also enhances computational capacity by deploying across multiple servers as the scale of hardware equipment increases.

4.1 Testbed Architecture

The large-capacity core switching equipment consists of a main control board, a service board, and a switching board. In this study, we develop vFabric on two servers. The main control board and the switching board are deployed on Server A, while the service board is deployed on Server B. In vFabric, the NP and switching processes emulate the functionality of the chips on actual network cards. Server A supports a maximum scale of 32 switching processes, while Server B supports up to 36 NP processes and 32 switching processes. Each NP process is connected to 20 port threads, each capable of sending and receiving packets at 100 Gbit/s. In vFabric, the packet transmission route is as follows: packets are generated from the ports and sequentially handled by the NP processes in Server A, the switching processes in Server A, the switching processes in Server B, the switching processes in Server A, the NP processes in Server A, and finally the port threads. All switching processes utilize round-robin scheduling to transmit the packets (Fig. 3).

The NP and switching processes have a similar architecture, consisting of a shell and a core (Fig. 4). The shell primarily implements data forwarding, while the core can execute packet processing modules for packet manipulation. In the

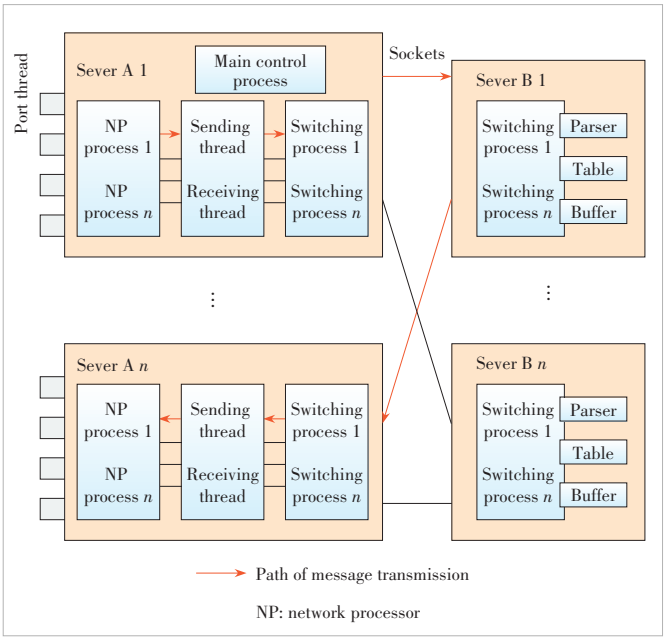


Figure 3. Framework of the digital twin model

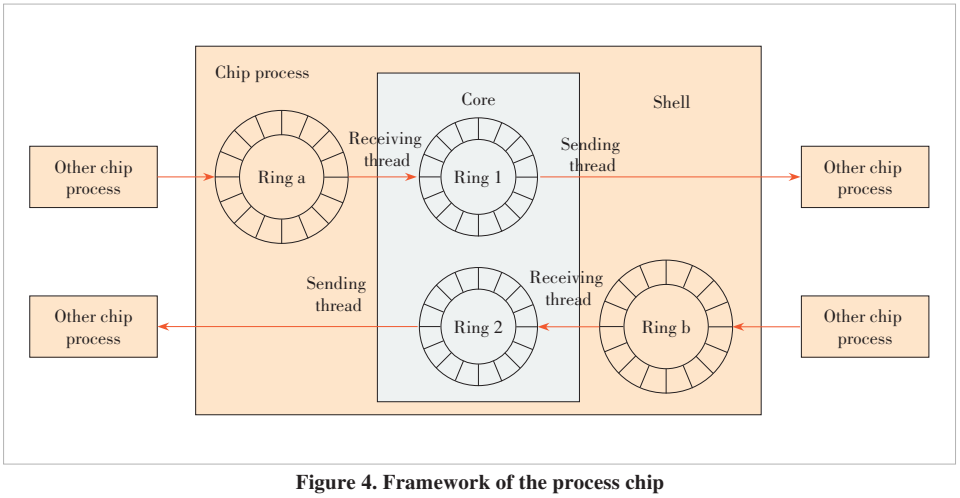


Figure 4. Framework of the process chip

shell, there are two queues (Ring a and Ring b) and two threads (receiving and sending threads). The core also contains two queues (Ring 1 and Ring 2). The receiving thread transfers packets between the shell and core through different queues, while the sending thread transfers packets to other chip processes.

4.2 Inter-Process Communication

The simulation process in vFabric involves a continuous flow of packet transmission and reception, which may impede the simulation speed and rapidly consume computation resources. To address this issue, we pre-configure memory based on simulation patterns and adopt a high-speed inter-process communication solution based on a multi-threading mechanism. This solution is designed to simulate the communication process between chips.

For small-scale deployments, vFabric can operate on a single server. We utilize the ring queue based on shared memory to facilitate communication between chip processes. The ring queue is a data structure that connects the front and rear ends circularly, following the first-in-first-out (FIFO) principle. It utilizes a linear array for storing data and offers simple data organization and efficient management.

In this paper, we present the implementation of high-speed data exchange between chip processes utilizing a producer-consumer model, as illustrated in Fig. 5. The model involves two threads: the producer and the consumer, which interact by reading from and writing to shared memory. The producer and consumer need to perform mutual exclusion operations to ensure data accuracy and safety.

For the ring queue, the sending thread acts as the producer, while the receiving thread serves as the consumer. Ensuring sole control over the circular queue is of utmost importance, permitting manipulation by a solitary thread exclusively at any given time. Specifically, when multiple producers write to the ring queue at the same time, only one thread is allowed to write to the queue. Consumers should also adhere to this principle in order to maintain data integrity and prevent conflicts.

On a large scale, vFabric is deployed across multiple servers, utilizing sockets for communication between multiple servers. The socket is a widely used network communication technology based on the Transmission Control Protocol/Internet Protocol (TCP/IP) protocol. It provides two endpoints for bidirectional host-to-host interaction. In modern networks, sockets have various applications and can facilitate data transmission, control, and management through various protocols, meeting the requirements of different scenarios.

4.3 Synchronization Mechanism

The vFabric platform utilizes a distributed simulation system across multiple servers. In a distributed system, time is a pivotal concept. Simultaneously, the chip processes also rely on time for synchronized interaction.

Currently, commonly used time synchronization methods in distributed simulation systems include the Network Time Protocol (NTP) and the Berkeley algorithm[38]. NTP is used to synchronize the time of various nodes in a computer network to ensure time consistency and accuracy. In contrast, the Berkeley algorithm achieves time synchronization in distributed systems by selecting a reference node to provide accurate time information and synchronizing the clocks of other nodes through communication with the reference node.

However, these techniques have limitations. NTP is not suitable for all types of distributed systems, and the Berkeley algorithm requires a central server, which may introduce a single point of failure. Therefore, we propose a time synchronization method for large-scale distributed simulation systems based on multi-level management (Fig. 6).

The time synchronization framework consists of three roles:

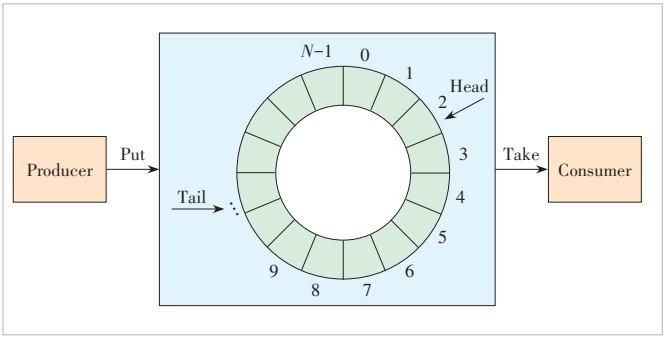


Figure 5. Framework of the producer-consumer model

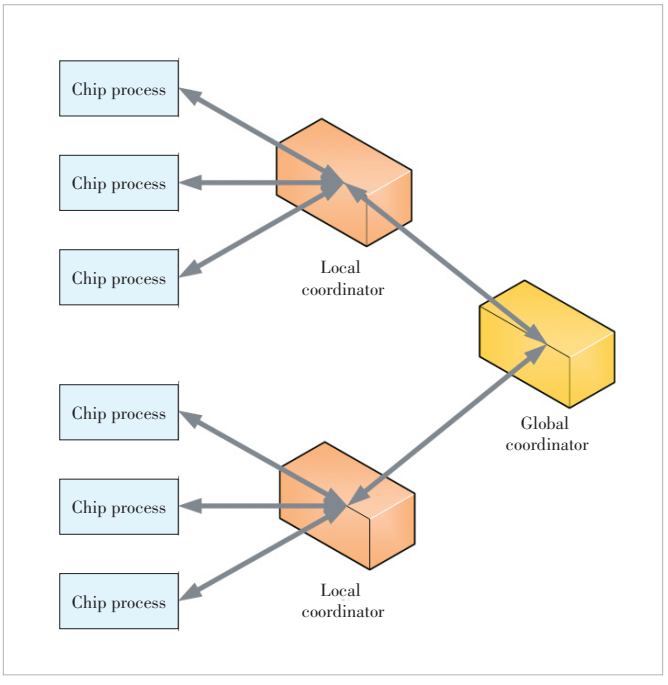


Figure 6. Framework of the synchronization module

the global coordinator, local coordinator, and member node. First, all member nodes send timestamps to the local coordinator node after each time slice and then enter a blocked state. The local coordinator node receives the timestamps from all the member nodes within its management range, sends the timestamps to the global coordinator node, and enters a blocked state. Upon receiving the timestamps from all local coordinator nodes, the global coordinator node advances the global clock and notifies all local coordinator nodes to continue execution. After receiving the response, the member nodes continue their computational tasks, completing time synchronization. This process of time synchronization among the three types of nodes continues iteratively until the simulation concludes.

4.4 Exception Handling

To emulate real-world scenarios involving the process of a chip being uploaded and offloaded due to various factors such as equipment failure, maintenance, and updates, vFabric incorporates periodic online and offline operations on the chip processes.

These operations often lead to fluctuations in system load, such as some nodes becoming overloaded while others remaining underutilized, which can impact the performance of the simulation system. To address such issues and ensure stability and efficiency in the cluster simulation system, effective handling of device online and offline processes is required to achieve traffic load balancing. Load balancing involves distributing the workload evenly across the nodes in the simulation cluster, thereby optimizing resource utilization and preventing any individual node from becoming overloaded.

To address the challenges, this paper proposes a dynamic traffic load-balancing algorithm based on real-time device status information (e.g., cache utilization and CPU occupancy). By leveraging the real-time status information, the algorithm intelligently selects appropriate forwarding paths and dynamically reschedules data packets based on the actual device conditions. When a chip comes online, the packets are rerouted among all the chips in the system (Fig. 7a). This ensures that the new chip can participate in the packet forwarding process. On the other hand, when a chip goes offline, the packets originally residing in that chip are evenly distributed to other available chips (Fig. 7b). This ensures maximum transmission throughput to avoid packet loss while maximizing system performance.

5 Implementation of VFabric

In this section, we present the implementation of vFabric in detail. We develop vFabric on the Linux platform through C++.

5.1 Implementation of Synchronization

In this paper, we establish a synchronization mechanism

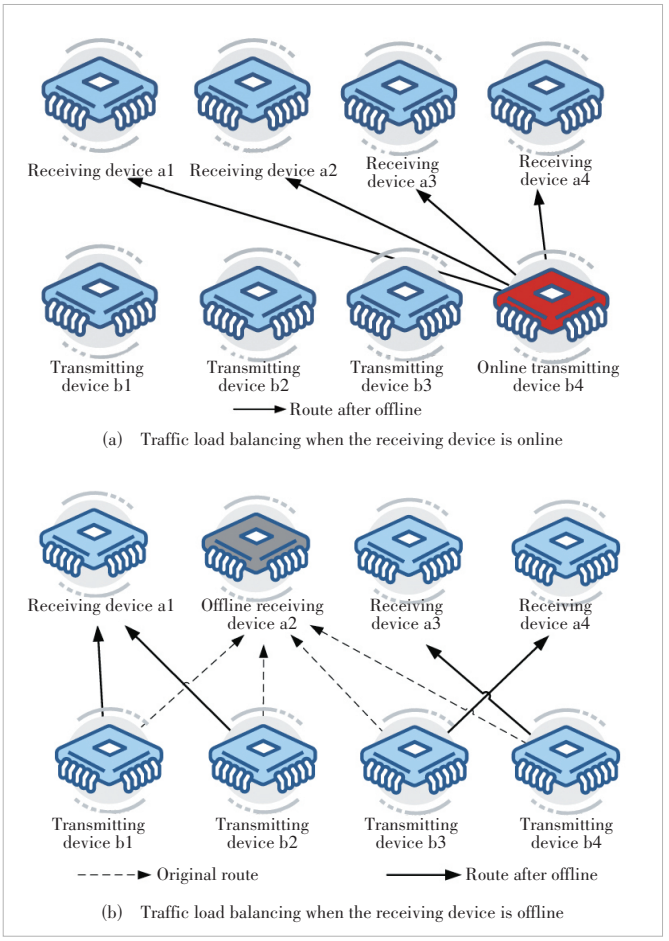
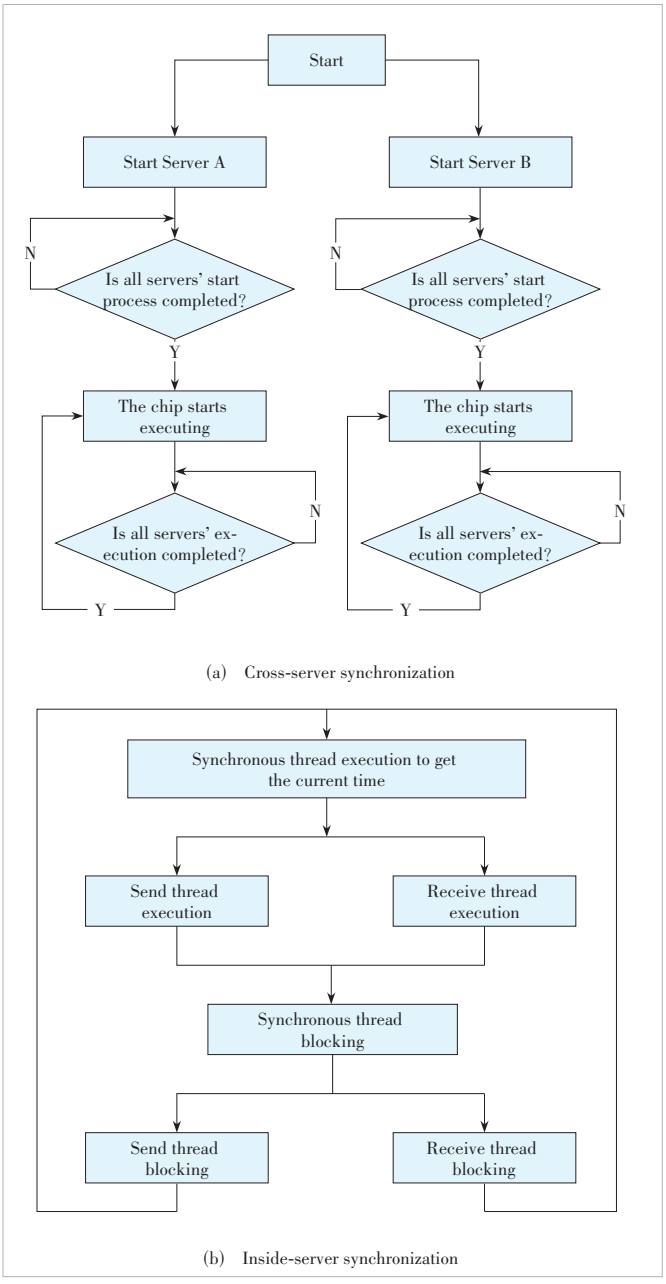


Figure 7. Schematic diagrams of online and offline processes

based on Redis and semaphores. Redis[39] is a mainstream non-relational database that is distributed and scalable, with high-performance. The semaphore, a mechanism for synchronous control among multiple threads or processes, coordinates the access sequence among different threads and processes to avoid data inconsistency. We deploy a Redis database on each server and connect them to form a distributed cluster. Each server can obtain the current time through the Redis database and perform corresponding time synchronization operations.

In the synchronization process between servers, each server contains a global synchronization thread. Each server is executed alternately with a global synchronization thread until the simulation concludes (Fig. 8a).

In the synchronization process within the server, the threads inside the chip (such as receiving and sending threads) and the synchronization thread alternately execute the P(wait) operation and V(signal) operation on the synchronization semaphore (Fig. 8b). P(wait) operation can block process execution, and V(signal) operation can resume process execution. With each alternation, the system time increases by one time slice until the simulation concludes.



(a) Cross-server synchronization

(b) Inside-server synchronization

Figure 8. Flow charts of synchronization

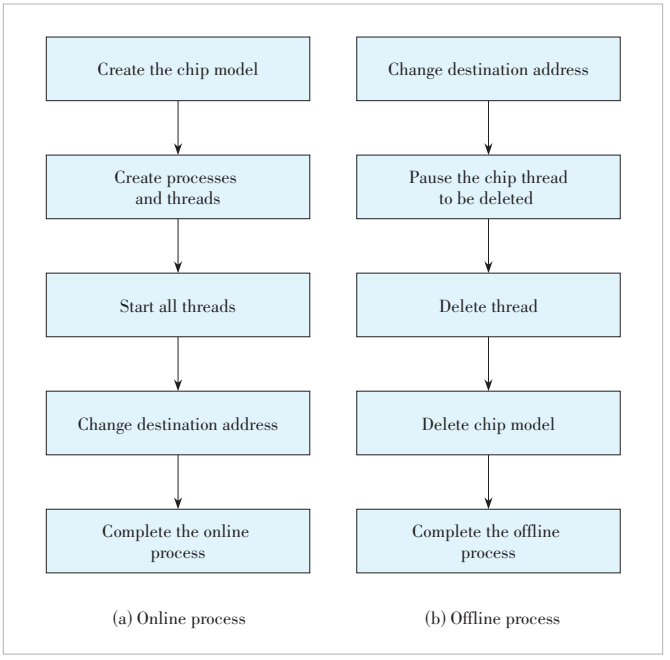


Figure 9. Flow charts of exception handling

5.2 Implementation of Exception Handling

During both online and offline processes of the chips, it is crucial to effectively handle related events such as routing planning, process management, process creation, and termination.

During the online process (Fig. 9a), the following steps are executed: 1) The required models (e.g., variables and pointers) are created; 2) once the models are created, the related processes and threads are initialized and blocked; 3) the synchronization thread starts alongside other chip threads; 4) the destination addresses of the packets are modified to ensure load balancing.

During the offline process (Fig. 9b), the following steps need to be taken: 1) The destination addresses of the packets are modified to prevent data loss; 2) while waiting for the synchronization thread to execute, the required chip threads are paused and removed; 3) the chip model is deleted to complete the device offline process.

The following is a typical process that a system performs over time. At the beginning of the offline process, the system time is recorded as t_1, and the destination addresses of the packets are immediately updated. If the destination chip of a packet has already gone offline, the destination address is changed to another available chip. After waiting until the system time reaches t_{1+nT}, the detection of destination addresses of internal chip packets is stopped. At this point, there are no more packets in the system with a destination address belonging to a chip that has already gone offline.

Here, t_1 is any time and T in t_{1+nT} represents the size of the time slice, and n is a variable that can be modified based on the server's performance. This variable can be determined according to specific circumstances to ensure that the detection and forwarding of internal chip packets are completed within an appropriate time frame, effectively preventing the processing of packets destined for offline chips.

6 Performance Evaluation

In this section, we conduct extensive experiments to evaluate the performance of vFabric and collect statistics as the scale gradually increases.

6.1 Experiment Settings

1) Platform: The evaluation platform is a workstation carrying an Intel(R) Xeon(R) Silver 4210R CPU (each has 80

cores). The RAM of the workstation is 260 GB and the operating system is Linux Ubuntu 20.10. The version of Redis is 6.0.8.

2) Data scale: As mentioned earlier, the servers are divided into two types (A and B). Server A can support a maximum of 36 NP processes and 32 switching processes, while Server B has a maximum of 32 switching processes. Each NP is connected to 20 100 Gbit/s port threads. We simulate the state of a real device forwarding 10 ms of traffic, while the total operation takes slightly longer than 10 ms. The additional time is allocated after 10 ms to ensure that all packets are fully transmitted. The delay depends on system characteristics and simulation requirements. Each port sends a packet of varying length every 25 ns. Table 1 summarizes the proportion of packets of different lengths. For example, the proportion of 64-byte messages is 449/1 000.

6.2 Results and Analysis

In the vFabric, each port generates 400 000 packets within a time interval of 10 ms. Table 2 summarizes the number of packets forwarded by different chips in a large-scale scenario.

The key indicators of the vFabric are the simulation time and packet loss rate. The simulation time of the system increases gradually with scale. For small-scale scenarios (4 NP processes and 32 switching processes), the total simulation time is approximately 100 s. For medium-scale scenarios (36 NP processes and 32 switching processes), the total simulation time is approximately 1 000 s. For large-scale scenarios (one Server A and one Server B), the total simulation time is approximately 4 800 s. These are generally in line with the expected requirements.

Different synchronization algorithms have been introduced above. We compare the time taken by two synchronization methods, the Berkeley algorithm and the multi-level management, as shown in Fig. 10. The multi-level management approach we adopted has a significant optimization effect. Specifically, the designed total simulation time is designed to be 10 ms, but in practice, it slightly exceeds this duration. Additional time is allocated after 10 ms to ensure all packets are fully transmitted. The length of this delay depends on the system characteristics and simulation requirements. Experimental results (Fig. 11) show that packet loss occurs when the simulation time is insufficient, and the packet loss rate gradually decreases as the simulation time increases. However, an excessively long simulation time leads to prolonged simulation duration. In the vFabric, the total simulation time is set to 10.1 ms.

Table 1. Proportion of packets of different lengths

Packet Length/B	Weight
64	449
130	160
260	200
577	80
1 518	110
9 000	1

Table 2. Number of packets forwarded by different chips

Type of Device	Number of Packets
Port	400 000
NP	8 000 000
Switch	9 000 000
Total server	28 800 000

NP: network processor

7 Conclusions

In this paper, we present a large-capacity core switching equipment digital twin platform. The simulation platform primarily consists of NP chips and switching chips, with a simu-

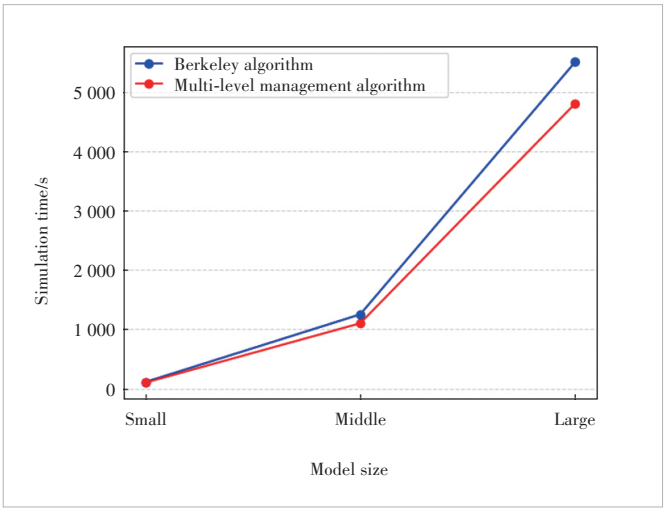


Figure 10. Simulation time of different synchronization algorithms

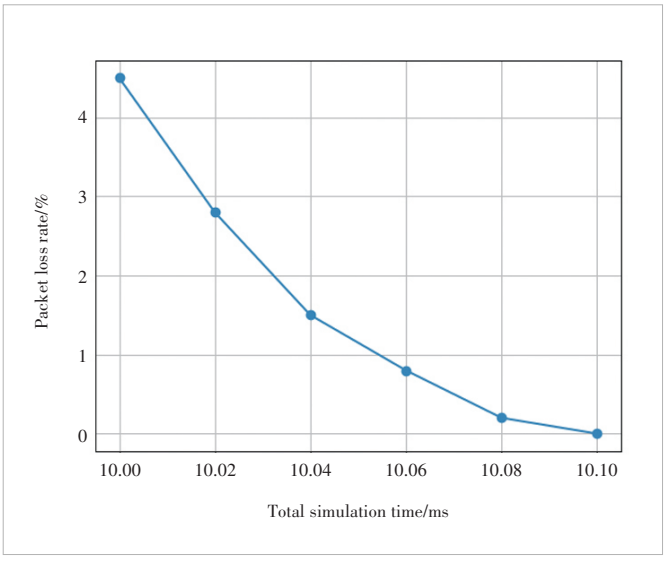


Figure 11. Change of the packet loss rate with time

lation time of 10 ms. The simulation results demonstrate the platform's efficient and reliable simulation capabilities. It accurately replicates the operational state of large-capacity core switching equipment. Moreover, we have successfully implemented time synchronization and the ability to dynamically bring chip processes online and offline, further enhancing the platform's functionality. The digital twin platform offers valuable applications in diagnosing and troubleshooting network failures. It assists engineers in promptly identifying and resolving issues, thereby enhancing the maintainability and manageability of the network. In the future, we will further enhance the platform to meet more complex and diverse simulation requirements.

References

[1] GÜNDOĞAN C, AMSÜSS C, SCHMIDT T C, et al. Content object security in the Internet of Things: challenges, prospects, and emerging solutions [J]. IEEE transactions on network and service management, 2022, 19(1): 538 – 553. DOI: 10.1109/TNSM.2021.3099902
[2] SHANG B D, LIU L J, MA J C, et al. Unmanned aerial vehicle meets vehicle-to-everything in secure communications [J]. IEEE communications magazine, 2019, 57(10): 98 – 103. DOI: 10.1109/MCOM.001.1900170
[3] YAN S C, ZHAO Z Y, GAO D Q, et al. Research on fuzzy location method of power communication network fault using digital twin model based on weight coefficient [C]//Proc. IEEE 6th Conference on Energy Internet and Energy System Integration. IEEE, 2022: 1779 – 1783. DOI: 10.1109/EI256261.2022.10116230
[4] CHEN M L, SHAO J, GUO S X, et al. Convoy_DTN: a security interaction engine design for digital twin network [C]//Proc. IEEE Globecom Workshops (GC Wkshps). IEEE, 2021. DOI: 10.1109/gcwkshps52748.2021.9682031
[5] WU Y W, ZHANG K, ZHANG Y. Digital twin networks: a survey [J]. IEEE Internet of Things journal, 2021, 8(18): 13789 – 13804. DOI: 10.1109/JIOT.2021.3079510
[6] RAZA M, KUMAR P M, HUNG D V, et al. A digital twin framework for industry 4.0 enabling next-gen manufacturing [C]//Proc. 9th International Conference on Industrial Technology and Management (ICITM). IEEE, 2020: 73 – 77. DOI: 10.1109/ICITM48982.2020.9080395
[7] LIU J, YANG H, NIU H C, et al. Digital twin civil aviation research airport for aircraft security and environment protection [C]//Proc. IEEE 4th International Conference on Civil Aviation Safety and Information Technology (ICCASIT). IEEE, 2022: 408 – 412. DOI: 10.1109/ICCASIT55263.2022.9986522
[8] SHRIVASTAVA M, CHUGH R, GOCHHAIT S, et al. A review on digital twin technology in healthcare [C]//Proc. International Conference on Innovative Data Communication Technologies and Application (ICIDCA). IEEE, 2023: 741 – 745. DOI: 10.1109/ICIDCA56705.2023.10099646
[9] CHEN J H, DENG R Q, GUO Y Y, et al. Research on network management technology of power line carrier communication in low-voltage distribution network based on digital twin [C]//Proc. 7th International Conference on Computer and Communications (ICCC). IEEE, 2021: 2112 – 2116. DOI: 10.1109/ICCC54389.2021.9674420
[10] RASHEED A, SAN O, KVAMSDAL T. Digital twin: values, challenges and enablers from a modeling perspective [J]. IEEE access, 2020, 8: 21980 – 22012. DOI: 10.1109/ACCESS.2020.2970143
[11] ZHAO Z Y, YAN S C, GAO D Q, et al. Research on digital twin network architecture for power grid telecommunication system [C]//Proc. IEEE 6th Conference on Energy Internet and Energy System Integration. IEEE, 2022: 1868 – 1874. DOI: 10.1109/EI256261.2022.10116995
[12] EL MARAI O, TALEB T, SONG J. Roads infrastructure digital twin: a step toward smarter cities realization [J]. IEEE network, 2021, 35(2): 136 – 143. DOI: 10.1109/MNET.011.2000398
[13] ZHU Y H, CHEN D Y, ZHOU C, et al. A knowledge graph based construction method for Digital Twin Network [C]//Proc. IEEE 1st International Conference on Digital Twins and Parallel Intelligence (DTPI). IEEE, 2021. DOI: 10.1109/dtpi52967.2021.9540177
[14] AUTIOSALO J, SIEGEL J, TAMMI K. Twinbase: open-source server software for the digital twin web [J]. IEEE access, 2021, 9: 140779 – 140798. DOI: 10.1109/ACCESS.2021.3119487
[15] KURUVATTI N P, HABIBI M A, PARTANI S, et al. Empowering 6G communication systems with digital twin technology: a comprehensive survey [J]. IEEE access, 2022, 10: 112158 – 112186. DOI: 10.1109/ACCESS.2022.3215493
[16] FU X Y, YUAN Q, LIU S F, et al. Communication-efficient decision-making of digital twin assisted Internet of vehicles: a hierarchical multi-agent reinforcement learning approach [J]. China communications, 2023, 20(3): 55 – 68. DOI: 10.23919/JCC.2023.03.005
[17] QIN W B, ZHANG C, YAO H P, et al. Stackelberg game-based offloading strategy for digital twin in Internet of vehicles [C]//Proc. International Wireless Communications and Mobile Computing (IWCMC). IEEE, 2023: 1365 – 1370. DOI: 10.1109/IWCMC58020.2023.10182450
[18] PILLAI R, BABBAR H. Digital twin for edge computing in smart vehicular systems [C]//Proc. International Conference on Advancement in Computation & Computer Technologies (InCACCT). IEEE, 2023: 1 – 5. DOI: 10.1109/InCACCT57535.2023.10141784
[19] DAI Y Y, ZHANG Y. Adaptive digital twin for vehicular edge computing and networks [J]. Journal of communications and information networks, 2022, 7(1): 48 – 59. DOI: 10.23919/JCIN.2022.9745481
[20] DAI C G, YANG K, DENG C J. A service placement algorithm based on merkle tree in MEC systems assisted by digital twin networks [C]//Proc. IEEE 21st International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS). IEEE, 2022: 37 – 43. DOI: 10.1109/IUCC-CIT-DSCI-SmartCNS57392.2022.00020
[21] GUO Y, ZHUANG Y, LI X, et al. Time-frequency correlated network sensing edge deployment for digital twin [C]//Proc. 4th International Academic Exchange Conference on Science and Technology Innovation (IAECST). IEEE, 2022: 1183 – 1187. DOI: 10.1109/IAECST57965.2022.10062032
[22] NJOKU J N, NKORO E C, MEDINA R M, et al. Leveraging digital twin technology for battery management: a case study review [J]. IEEE access, 2022, 13: 21382 – 21412, DOI: 10.1109/ACCESS.2025.3531833
[23] TAO Y H, WU J, LIN X, et al. DRL-driven digital twin function virtualization for adaptive service response in 6G networks [J]. IEEE networking letters, 2023, 5(2): 125 – 129. DOI: 10.1109/LNET.2023.3269766
[24] LU Y L, MAHARJAN S, ZHANG Y. Adaptive edge association for wireless digital twin networks in 6G [J]. IEEE Internet of Things journal, 2021, 8(22): 16219 – 16230. DOI: 10.1109/JIOT.2021.3098508
[25] WEI Z Y, WANG S T, LI D, et al. Data-driven routing: a typical application of digital twin network [C]//Proc. IEEE 1st International Conference on Digital Twins and Parallel Intelligence (DTPI). IEEE, 2021. DOI: 10.1109/dtpi52967.2021.9540073
[26] RAJ D R R, SHAIK T A, HIRWE A, et al. Building a digital twin network of SDN using knowledge graphs [J]. IEEE access, 2023, 11: 63092 – 63106. DOI: 10.1109/ACCESS.2023.3288813
[27] ONO S, YAMAZAKI T, MIYOSHI T, et al. AMoND: area-controlled mobile ad-hoc networking with digital twin [J]. IEEE access, 2023, 11: 85224 – 85236. DOI: 10.1109/ACCESS.2023.3304374
[28] Huawei. NetEngine 8000 series router [EB/OL]. [2023-06-15]. https://e.huawei.com/cn/products/routers/ne8000
[29] Cisco. NAT order of operation [EB/OL]. [2023-06-15]. https://www.cisco.com/c/zhcn/products/routers/8000-series-routers/index.html

[30] ZTE. ZXR10 6800 multi-service router [EB/OL]. [2023-06-15]. https://www.zte.com.cn/china/products/bearer/router/ZXR10-6800.html
[31] ZTE. ZXR10 8900E series core switch [EB/OL]. [2023-06-15]. https://www.zte.com.cn/china/products/bearer/Ethernet-Switch/8900E-CH.html
[32] JIN S, ZHANG Z B, CHAKRABARTY K, et al. Hierarchical symbol-based health-status analysis using time-series data in a core router system [J]. IEEE transactions on computer-aided design of integrated circuits and systems, 2020, 39(3): 700 – 713. DOI: 10.1109/TCAD.2018.2890681
[33] JIN S, ZHANG Z B, CHAKRABARTY K, et al. Self-learning health-status analysis for a core router system [C]//Proc. IEEE International Test Conference (ITC). IEEE, 2018: 1 – 10. DOI: 10.1109/TEST.2018.8624712
[34] WU C C, QIAO L F, CHEN Q H. Design of a 640-gbps two-stage switch fabric for satellite on-board switches [J]. IEEE access, 2020, 8: 68725 – 68735. DOI: 10.1109/ACCESS.2020.2986300
[35] GLABOWSKI M, LEITGEB E, SOBIERAJ M, et al. Analytical modeling of switching fabrics of elastic optical networks [J]. IEEE access, 2020, 8: 193462 – 193477. DOI: 10.1109/ACCESS.2020.3033186
[36] MADUREIRA A L R, ARAÚJO F R C, ARAÚJO G B, et al. NDN fabric: where the software-defined networking meets the content-centric model [J]. IEEE transactions on network and service management, 2021, 18(1): 374 – 387. DOI: 10.1109/TNSM.2020.3044038
[37] ALMASAN P, FERRIOL-GALMES M, PAILLISSE J, et al. Network digital twin: context, enabling technologies, and opportunities [J]. IEEE communications magazine, 2022, 60(11): 22 – 27. DOI: 10.1109/mcom.001.2200012
[38] FANG Z H, GAO Y. Delay compensated one-way time synchronization in distributed wireless sensor networks [J]. IEEE wireless communications letters, 2022, 11(10): 2021 – 2025. DOI: 10.1109/LWC.2022.3189744
[39] BEN SEGHIER N, KAZAR O. Performance benchmarking and comparison of NoSQL databases: Redis vs MongoDB vs Cassandra using YCSB tool [C]//Proc. International Conference on Recent Advances in Mathematics and Informatics (ICRAMI). IEEE, 2021. DOI: 10.1109/icrami52622.2021.9585956

Biographies

WANG Qianglin received his MS degree in electronic information from University of Electronic Science and Technology of China in 2024. He joined Alibaba Corporation in 2024 and is now a network engineer there.

ZHANG Xiaoning (xnzhang@uestc.edu.cn) received his MS and PhD degrees in electronic information from the School of Communication and Information Engineering, University of Electronic Science and Technology of China (UESTC) in 2005 and 2007, respectively. Since 2007, he has been a professor with UESTC. His current research interests include software-defined networking and virtualization of network functions.

YANG Yi received his MS degree in electronic information from University of Electronic Science and Technology of China in 2024. He joined Tianyi Cloud Corporation in 2024.

FAN Chenyu received her MS degree in electronic information from University of Electronic Science and Technology of China in 2024. She joined EricssonCloud Corporation in 2024.

YUE Yangyang is a senior engineer at ZTE Corporation. She is engaged in application software development for wired communications.

WU Wei is an engineer at ZTE Corporation. He is engaged in application software development for wired communications. His primary research focuses on large-scale router penetration testing, performance simulation, and functional development.

DUAN Wei is a senior R&D chief engineer at ZTE Corporation. He is engaged in the key technology research of IP networks and intelligent computing center networks. He has applied for more than 30 patents.

Precise Location of Passive Intermodulation in Long Cables by Fractional Frequency Based Multi-Range Rulers

DONG Anhua[1], LIANG Haodong[2], ZHU Shaohao[2], ZHANG Qi[1], ZHAO Deshuang[1]
(1. University of Electronic Science and Technology of China, Chengdu 611731, China;
2. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202501013
https://kns.cnki.net/kcms/detail/34.1294.TN.20250218.1313.002.html, published online February 18, 2025
Manuscript received: 2024-03-04

Abstract: A novel method is developed by utilizing the fractional frequency based multi-range rulers to precisely position the passive intermodulation (PIM) sources within radio frequency (RF) cables. The proposed method employs a set of fractional frequencies to create multiple measuring rulers with different metric ranges to determine the values of the tens, ones, tenths, and hundredths digits of the distance. Among these rulers, the one with the lowest frequency determines the maximum metric range, while the one with the highest frequency decides the highest achievable accuracy of the position system. For all rulers, the metric accuracy is uniquely determined by the phase accuracy of the detected PIM signals. With the all-phase Fourier transform method, the phases of the PIM signals at all fractional frequencies maintain almost the same accuracy, approximately 1°(about 1/360 wavelength in the positioning accuracy) at the signal-to-noise ratio (SNR) of 10 dB. Numerical simulations verify the effectiveness of the proposed method, improving the positioning accuracy of the cable PIM up to a millimeter level with the highest fractional frequency operating at 200 MHz.

Keywords: passive intermodulation; location; multi-range

Citation (Format 1): DONG A H, LIANG H D, ZHU S H, et al. Precise location of passive intermodulation in long cables by fractional frequency based multi-range rulers [J]. ZTE Communications, 2025, 23(1): 101 - 106. DOI: 10.12142/ZTECOM.202501013
Citation (Format 2): A. H. Dong, H. D. Liang, S. H. Zhu, et al., "Precise location of passive intermodulation in long cables by fractional frequency based multi-range rulers," *ZTE Communications*, vol. 23, no. 1, pp. 101 - 106, Mar. 2025. doi: 10.12142/ZTECOM.202501013.

1 Introduction

Passive intermodulation (PIM) interference has become increasingly prominent with the growing demands for high-power, wideband, and multi-carrier microwave communication systems, such as high-speed 5G and 6G wireless communications, indoor distributed antenna systems, and satellite communications[1]. PIM refers to the nonlinear effect in high-power passive microwave devices due to the coupling mechanism of the electro-thermal and multi-physical fields. PIMs are generally generated as distorted products from the emitting of high-power signals or multi-carrier networks, which can interfere with the whole communication process and eventually weaken the performance of the communication systems[2]. For a communication system, the PIM level has become an important technical index to evaluate the performance. Therefore, the strict PIM level is expected to minimize the interference and improve the system capacity. To obtain the limited PIM level and reduce the PIM interference efficiently, the generation mechanism of PIM interference has been first investigated. The results obtained in Refs. [3 - 6] indicate that the PIM interference is usually generated by multiple physical factors causing shape alterations and imperfect connection, such as temperature and humidity, oxidation and pollution of clean surfaces, and loose connection of devices. PIM interference can be produced at the formed unknown sources that are formed within the radio frequency (RF) cables. Therefore, to further reduce PIM interference from the unknown PIM sources existing in the RF cables, the precise detection and location of PIM sources has attracted much interests.

Recently, the near-field scanning method has been used to detect the non-enclosed PIM source[7]. Based on the field nephogram of the plane above the device under test (DUT), which is constructed using the measured amplitude and phase information of the magnetic field, the field nephograms of the plane below the DUT are estimated. From the estimated nephograms, the positions of the PIM sources can be located. Similarly, emission source microscopy (ESM) has been developed to locate PIM sources by measuring the amplitude and phase of the field on a plane a few wavelengths away from the DUT[8]. Different

from the amplitude and phase information evaluated in Refs. [7] and [8], the position of PIM sources in the base station antenna is identified by the acoustic vibration method[9], which detects the intensity of the modulated PIM signal. With the measured signals, another interesting method of the K-space multi-carrier signals is proposed to locate multiple PIM sources in microwave systems[10]. However, the relatively low positioning accuracy of these methods needs to be improved.

In this paper, a high positioning accuracy method called fractional frequency based multi-range rulers (FF-MRR) has been developed to locate the PIM sources in RF cables precisely. The range of the PIM sources is obtained by processing each ranging datum of each fractional frequency signal. The diverse frequencies of ruler signals can be widely used across different scenarios without constructing complicated systems. The ruler signal is obtained by mixing a group of signals and the local oscillator signal in batches. Additionally, the proposed method is not limited by the narrow-band bandwidth because the fractional frequency signals can still be obtained by adjusting the local oscillator signal. With the adopted fractional frequencies, the higher frequency signals and the lower ones in the multi-range rulers can guarantee a high positioning accuracy and a long measured distance.

The remainder of this paper is as follows. The system model of FF-MRR for positioning the PIM sources is proposed in Section 2. The calculations of the precise location of PIM by FF-MRR are deduced in detail in Section 3. In Section 4, the numerical simulation results of locating the PIM sources by the proposed FF-MRR method are discussed and analyzed. Section 5 concludes the paper.

2 System Model of FF-MRR

Compared with the pulse method based on timing ranging, it is easier for the phase method based on phase ranging to achieve higher accuracy[11]. However, the periodic ambiguity of phases makes the phase based methods difficult to precisely position PIM, when the cable length exceeds one wavelength of the detected signal. Fig. 1 shows the schematic diagram of a single metric ruler based locating system using the phase method, which calculates the location of the PIM source by multiplying half of the wavelength by the ratio obtained from dividing 2π by the measured phase. Based on such a metric ruler, the maximum measurement distance is limited to half of the signal wavelength. Although one can lower the signal frequency to increase the maximum measurement distance, the position accuracy will decrease.

To achieve both a long measurement distance and high positioning accuracy, we have developed a novel method called FF-MRR. In our proposed method, three fractional frequencies are employed to build multiple metric rulers with different measurement ranges. The metric ruler based on the lowest frequency is to obtain the maximum measurement range while the metric ruler based on the highest frequency is to increase the positioning accuracy. The metric ruler based on the middle frequency is designed to eliminate the periodic ambiguity of phases. In this study, we call them the long ruler, fine ruler, and short ruler.

Fig. 2 illustrates the proposed FF-MRR system and its components for locating the PIM source. In the system, a fractional frequency generator is employed to generate signals operating at specific frequencies. During the positioning process, the fractional signal is modulated onto the first channel signal, which will carry the useful message. After passing through the filter and being amplified by the power amplifier, the signal is combined with the amplified signal from the second channel by the combiner to form a double-tone signal. To obtain the reference PIM signal and the real PIM signal from the RF cable, a forward coupler and a backward coupler are inserted between the combiner and the RF cable. The forward coupler extracts a small amount of power from the double-tone signals to generate a reference PIM signal by a passive mixer. The backward coupler captures the PIM signals generated by the RF cable. For both the reference and real PIM signals, a down-converter is utilized to lower the frequency of the PIM signals so that we can apply a low-rate analog-to-digital converter (ADC) to collect the PIM signals. Such a design does not directly sample the PIM signals at very high frequencies, thereby reducing the cost and complexity of the positioning system. After ADC captures the real PIM signals and the reference, MCU will perform the program to extract the phase of the PIM signal operating at the fractional frequency. For each fractional frequency, the system performs the phase detection for the PIM signal.

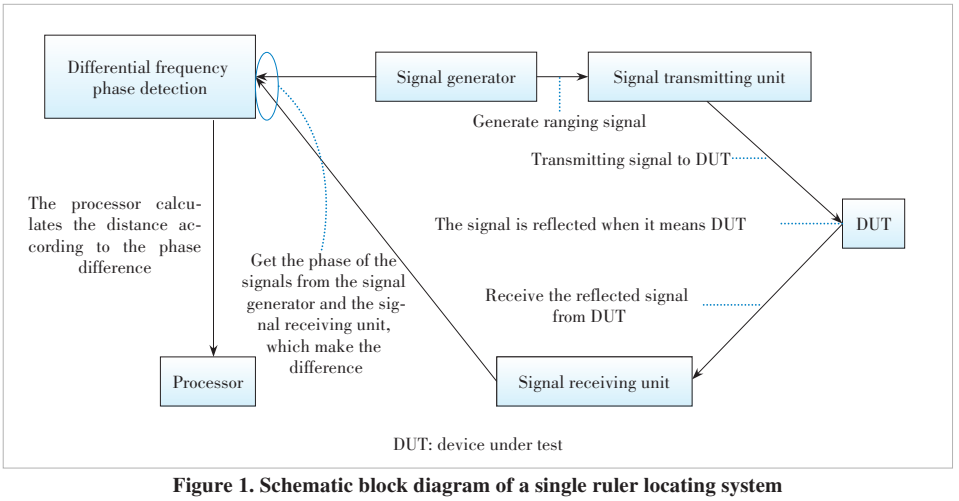


Figure 1. Schematic block diagram of a single ruler locating system

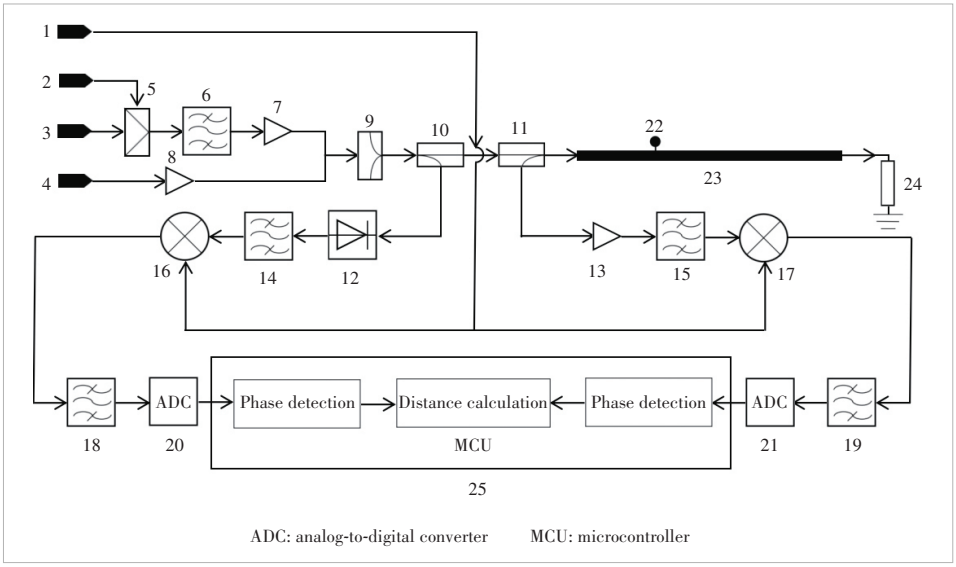


Figure 2. Block diagram of the FF-MRR system to locate PIM source. 1: local oscillator with f_{lo}, 2: fractional frequency control signal with f_0, 3: the first signal source at f_1, 4: the second signal source at f_2, 5: modulator, 6: filter, 7 and 8: power amplifiers, 9: combiner, 10: forward coupler, 11: backward coupler, 12: passive mixer, 13: low noise amplifier, 14 and 15: filters, 16 and 17: down converters, 18 and 19: low pass filters, 20 and 21: analog-to-digital converters, 22: PIM source, 23: RF cable, 24: matching load, and 25: MCU

3 PIM Location with FF-MRR

3.1 Range Distance Based on Phase Differences

Assume that the frequencies of double-tone signals passing through the combination of two channels are f_1 and f_2, and their initial phases are φ_1 and φ_2, respectively. Moreover, the measured distance of the double-tone signals D_1 can be obtained by $f_S + f_s$:

$$D_1 = \frac{c}{2f_1}\frac{\varphi_1}{2\pi} = \frac{c}{2f_2}\frac{\varphi_2}{2\pi} \tag{1},$$

where c is the velocity of the electromagnetic wave. From Eq. (1), $\varphi_1 = \frac{4f_1\pi D_1}{c}$ and $\varphi_2 = \frac{4f_2\pi D_1}{c}$ can be acquired. The third-order intermodulation signal is generated by the two signals with f_1 and f_2, whose frequency f_s is assumed as $f_s = 2f_1 - f_2$ with the initial phase $\varphi_s = 2\varphi_a - \varphi_b$. Similarly, the measured distance of the generated third-order intermodulation signal D_2 with the frequency f_s is formulated as:

$$D_2 = \frac{c}{2f_s}\frac{\varphi_s}{2\pi} = \frac{c}{2(2f_1 - f_2)}\frac{2\varphi_1 - \varphi_2}{2\pi} \tag{2}.$$

By substituting $\varphi_1 = \frac{4f_1\pi D_1}{c}$ and $\varphi_2 = \frac{4f_2\pi D_1}{c}$ into Eq. (2), it can be reformulated as:

$$D_2 = \frac{c}{2(2f_1 - f_2)}\frac{4\pi D_1}{c}\frac{(2f_1 - f_2)}{2\pi} = D_1 \tag{3}.$$

Eqs. (2) and (3) show that the measured distance depends on the phase differences caused by the transmission path.

3.2 Range Distance with FF-MRR

To locate the positions of the PIM sources, the phase difference from the third-order intermodulation signal caused by the transmission path should be first obtained. The referred PIM signal f_{IM3} generated by the double-tone signal with f_1 and f_2 is output from the forward coupler, which can be formulated as:

$$f_{IM3} = A_{IM3}\cos[2\pi(2f_1 - f_2 - 2f_0)t + (2\varphi_1 - \varphi_2 - 2\varphi_0)] \tag{4},$$

where A_{IM3} is the amplitude of the referred PIM signal, and f_0 and φ_0 are the frequency and initial phase of the ruler control signal.

The double-tone signals are input into the cable to be tested and transmitted to the position of the PIM source within Δt. Then, the double-tone signal undergoes nonlinear intermodulation, resulting in an intermodulation signal with the same frequency as the referred third-order intermodulation signal f_{IM3}. With the phase differences caused by the transmission path, the reflected intermodulation signal f'_{IM3} at the input port is received, which can be obtained by:

$$f'_{IM3} = A_{IM3}\cos[2\pi(2f_1 - f_2 - 2f_0)t + 2(\varphi_1 - \varphi_0) - \varphi_2 + 2\cdot 2\pi(2f_1 - f_2 - 2f_0)\Delta t] \tag{5}.$$

The phase differences φ_{IM3} caused by the transmission path can be obtained by:

$$\varphi_{IM3} = 2\pi(2f_1 - f_2 - 2f_0)(2\Delta t) \tag{6}.$$

Based on the measured phase discrimination accuracy of 1°, the ranges of wavelengths of the long ruler, fine ruler, and short ruler are $200\text{ m} \leqslant \lambda_L \leqslant 360\text{ m}$, $2\text{ m} \leqslant \lambda_F \leqslant 36\text{ m}$, and $0.2\text{ m} \leqslant \lambda_S \leqslant 3.6\text{ m}$, respectively. Correspondingly, the ranges of frequencies of the long ruler, fine ruler, and short ruler are $0.42\text{ MHz} \leqslant f_L \leqslant 1.5\text{ MHz}$, $4.2\text{ MHz} \leqslant f_F \leqslant 150\text{ MHz}$ and $42\text{ MHz} \leqslant f_S \leqslant 1500\text{ MHz}$, respectively. From the long ruler, the coarse distance D_L can be obtained by:

$$D_L = c\frac{\varphi_L}{4\pi f_L} \tag{7},$$

where $\varphi_L = \varphi_{IM3}$ is measured by MCU in Fig. 2. From the fine

ruler, the relative accurate range distance D_F can be obtained by:

$$\begin{cases} D_F = c\dfrac{\varphi_F + 2\pi K_F}{4\pi f_F} \\ K_F = \left[\dfrac{D_L - c\varphi_F/4\pi f_F}{\lambda_F}\right]_{\text{int}} \end{cases} \quad (8),$$

where φ_F is measured by MCU in Fig. 2, K_F is an integer from the fine ruler, and $[\ ]_{\text{int}}$ is the integer operator. More accurately, from the short ruler, the precise range distance D_S can be obtained by:

$$\begin{cases} D_S = c\dfrac{\varphi_S + 2\pi K_S}{4\pi f_S} \\ K_S = \left[\dfrac{D_F - c\varphi_S/4\pi f_S}{\lambda_S}\right]_{\text{int}} \end{cases} \quad (9),$$

where φ_S is also measured by MCU in Fig. 2, and K_S is an integer from the short ruler. Finally, the final measured distance of the PIM source D_{IM3} can be obtained from the most precise range distance, which can be formulated as

$$D_{\text{IM3}} = D_S \quad (10).$$

4 Simulation Results of PIM Location

4.1 Simulation Setup

The specific simulation conditions are set as shown in Tables 1 and 2.

According to the project requirements, the specified transmission frequency band ranges from 1 805 MHz to 1 880 MHz. Consequently, the setup of f_1 and f_2 has adopted two frequencies in the transmission band, namely 1 820 MHz and 1 880 MHz. In fact, this system is basically not limited by the frequency band and bandwidth. By adjusting the local oscillator signal f_{lo}, this system can be adaptable to frequency bands and bandwidths under various conditions. The positioning accuracy of PIM through FF-MMR depends on the highest fractional frequency in the fractional frequency based multi-range rulers. By adjusting the fractional frequency control signal with f_0, it is easy to reach the fractional frequency up to 200 MHz or even higher. Therefore, a conservative setup of 200 MHz is adopted for f_s here. f_0 is calculated based on the configured values of f_1, f_2, and f_s. The main function of the local oscillator signal f_{lo} is to down-convert the reference signal and the actual PIM signal so that we can use a lower rate ADC to collect PIM signals. The project requires a receiver sampling rate of 92.16 MHz. To make the sampling frequency close to 10 times the signal frequency, the local oscillator signal f_{lo} is set to 190 MHz, so that the signal at the receiving end can be reduced to 10 MHz. The local oscillator signal f_{lo} can be changed as needed based on actual requirements. Commonly seen on the market, the dielectric constant of radio frequency coaxial lines with foamed polyethylene (PE) as the dielectric layer is approximately between 1.4 and 2.0, while the ones with PE as the dielectric layer have a dielectric constant of around 2.3. Here, the relative dielectric constant of the cable is set to 2. For high-frequency signal transmissions, the signal-to-noise ratio usually needs to reach 15 dB or above. Here, a relatively poor communication environment (the SNR is assumed as 10 dB) has been chosen for simulation. Existing phase detection technologies generally can achieve phase detection accuracy in the millidegree range or higher. Here, a relatively conservative configuration of 1 degree is adopted for the phase detection accuracy.

Table 1. Setup of frequency-related conditions in the simulation

First Signal Source f_1/MHz	Second Signal Source f_2/MHz	Highest Fractional Frequency f_s/MHz	Fractional Frequency Control Signal f_0/MHz	Local Oscillator Signal f_{lo}/MHz
1 820	1 880	200	780	190

Table 2. Setup of other conditions in the simulation

Relative Dielectric Constant of Cable	Signal-to-Noise Ratio/dB	Phase Discrimination Accuracy/(°)
2	10	1

4.2 Numerical Results

Table 3 shows the range distances for the PIM source by FF-MRR considering the highest fraction frequency f_S. Generally, the distances of PIM sources range from 0.1 m to 20 m. The columns "Error #1" to "Error #5" represent the average of ten independent and repeatable trials. A total of 14 distance ranging cases from PIM sources is conducted using FF-MRR, and an average ranging error of 0.490 mm is obtained, which

Table 3. Simulation ranging error of 200 MHz ruler signal under different distances to be measured

Distance/m	Error#1/mm	Error#2/mm	Error#3/mm	Error#4/mm	Error#5/mm	Average Error/mm
20	0.692	0.795	0.751	0.780	0.721	0.747 8
18	0.467	0.557	0.287	0.287	0.467	0.413 0
16	0.336	0.336	0.236	0.336	0.236	0.296 0
14	0.741	0.719	0.732	0.727	0.732	0.730 2
12	0.300	0.519	0.339	0.191	0.300	0.329 8
10	0.332	0.575	0.413	0.494	0.332	0.429 2
7	0.433	0.357	0.509	0.509	0.585	0.478 6
5	0.280	0.280	0.508	0.166	0.166	0.280 0
3	0.689	0.737	0.737	0.717	0.746	0.735 2
1	0.504	0.262	0.357	0.452	0.357	0.386 4
0.7	0.259	0.259	0.450	0.641	0.641	0.450 0
0.5	0.658	0.710	0.710	0.658	0.658	0.678 8
0.3	0.780	0.693	0.564	0.607	0.607	0.650 5
0.1	0.174	0.399	0.286	0.174	0.286	0.263 8

is less than 1 mm. Furthermore, Table 4 provides a comparative analysis of the positioning accuracy of various technologies used for locating PIM sources. The proposed FF-MRR method achieves a positioning accuracy of approximately 1 mm, outperforming near-field scanning (10 mm), acoustic vibration (10 mm), K-space (37.5 mm) and ESM (5 mm). Consequently, the favorable errors of range distances indicate the precise location of PIM sources can be achieved by FF-MRR compared with the other positioning technologies.

4.3 Errors Analysis

Moreover, the SNR and transmission velocity from filters, mixers and other components affecting the positioning error-should be analyzed in the real measured environment.

Firstly, noise interference is considered in the positioning error, which is referred to as random Gaussian white noise. Fig. 3 shows the positioning error obtained by FF-MRR varies with the increased SNR. The maximum fractional frequencies of signals are considered as 7.5 MHz, 15 MHz, 30 MHz, 50 MHz, 100 MHz, 150 MHz and 200 MHz with the initial phase of 60 degrees. Additionally, the sampling frequency is 500 MHz. It can be found that a lower frequency results in a higher ranging error, because the ranging error is determined by the signal wavelength and the accuracy of the identification phase. When the accuracy of the identification phase remains unchanged, the longer the wavelength of the ranging signal, the greater the range error. Generally, as the SNR increases, the ranging errors decrease. The ranging error is declined to approximate 0.01 wavelengths when the SNR exceeds 10 dB with the frequency of ranging signal higher than 50 MHz.

Besides, the relative dielectric constant of transmission line can alter the velocity of electromagnetic wave, thereby influencing the ranging accuracy. Fig. 4 illustrates that the range errors vary with the relative dielectric constant of the transmission line. As the relative dielectric constant increases, the velocity of the electromagnetic wave decreases and the wavelength declines. Eventually, the ranging errors are reduced as well.

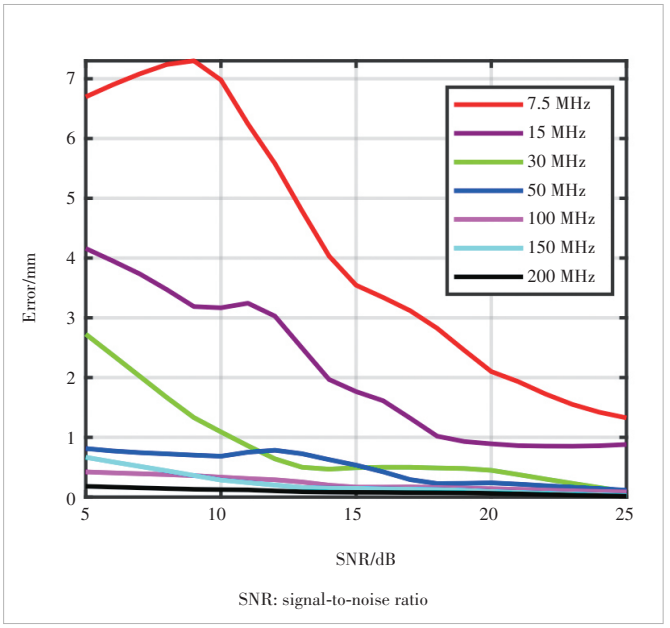


Figure 3. Positioning error versus SNR

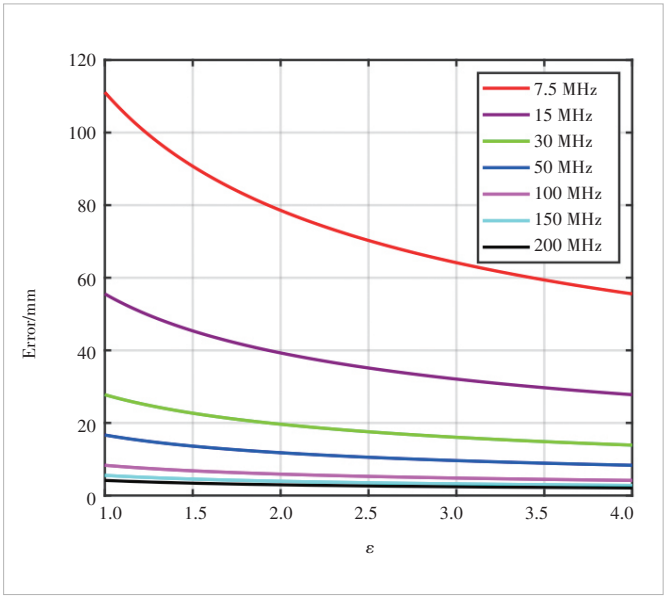


Figure 4. Influence of the relative dielectric constant of the RF cable on the positioning error at different fractional frequencies

5 Conclusions

In this paper, an FF-MRR method is proposed to locate the PIM sources in cables. In the FF-MRR method, fractional frequency signals across multiple ranges can be obtained. Higher frequencies enable high-positioning accuracy, while lower frequencies facilitate long range detection. Systematic simulations verify that the FF-MMR method has the advantage of high accu-

Table 4. Comparison of various PIM locating technologies

PIM Locating Technology	Scenario	Distance/m	Working Frequency/MHz	Error/mm
The near-field scanning[1, 12]	Microstrip line	0.21	935 – 960	10
Acoustic vibration[1, 9]	Antenna	–	1 850 – 1 990	10
K-space muti-carrier signals[10]	Cables	2.437	1 125 – 1 175	37.5
Emission source microsopy[8]	PCB	0.7	1 932 – 1 985	5
Our work	Cables	20	1 805 – 1 880	0.519

PCB: printed circuit board PIM: passive intermoludation

racy. When the highest frequency is 200 MHz, the positioning accuracy can reach the millimeter level. Meanwhile, it only requires flexible adjustment of the wavelengths of the ruler signal to locate PIM sources with FF-MRR, which just needs a few seconds in the whole locating process, highlighting the high efficiency advantage of the method. In addition, the PIM source location approach using the FF-MRR method and the proposed system diagram enploys a lower rate ADC without requiring other expensive instruments like scanning probes in near-field scanning methods. Therefore, this method also has the merits of lower cost and ease of portability. This method significantly improves the positioning performance in the PIM location technology for cable test scenarios compared with other methods. Nevertheless, the applicability of this method is somewhat restricted. It cannot be used in test scenarios involving antennas, printed circuit boards (PCBs), and the like.

References

[1] CAI Z H, LIU L, DE PAULIS F, et al. Passive intermodulation measurement: challenges and solutions [J]. Engineering, 2022, 14: 181 – 191. DOI: 10.1016/j.eng.2022.02.012
[2] ZHANG L, WANG H G, HE S T, et al. A segmented polynomial model to evaluate passive intermodulation products from low-order PIM measurements [J]. IEEE microwave and wireless components letters, 2019, 29(1): 14 – 16. DOI: 10.1109/LMWC.2018.2883719
[3] WANG X L, CHEN X, SUN D Q. A compact contactless waveguide bandpass filter for high sensitivity passive intermodulation measurement [C]// Proceedings of IEEE MTT-S International Wireless Symposium (IWS). IEEE, 2023: 1 – 3. DOI: 10.1109/IWS58240.2023.10223117
[4] ISHIBASHI D, KUGA N. Numerical analysis of DUT-size effect on PIM measurement using standing-wave coaxial tube [C]//Asia Pacific Microwave Conference. IEEE, 2009: 2609 – 2612. DOI: 10.1109/APMC.2009.5385244
[5] CANTALI G, DENIZ E, OZAY O, et al. PIM detection in wireless networks as an anomaly detection problem [C]//International Balkan Conference on Communications and Networking (BalkanCom). IEEE, 2023: 1 – 6. DOI: 10.1109/BalkanCom58402.2023.10167980
[6] WANG W B, WANG Y M, ZANG W X, et al. Physical mechanisms of passive intermodulation: a short review [C]//International Applied Computational Electromagnetics Society Symposium (ACES-China). IEEE, 2022: 1 – 3. DOI: 10.1109/ACES-China56081.2022.10064913
[7] XU Z. Research on electromagnetic interference source location algorithm based on near-field scanning [D]. Hangzhou: Zhejiang University, 2022. DOI: 10.27461/d.cnki.gzjdx.2022.000065
[8] YONG S, YANG S, ZHANG L, et al. Passive intermodulation source localization based on emission source microscopy [J]. IEEE transactions on electromagnetic compatibility, 2020, 62(1): 266 – 271. DOI: 10.1109/TEMC.2019.2938634
[9] YANG S, WU W, XU S, et al. A passive intermodulation source identification measurement system using a vibration modulation method [J]. IEEE transactions on electromagnetic compatibility, 2017, 59(6): 1677 – 1684. DOI: 10.1109/TEMC.2017.2705114
[10] ZHANG M, ZHENG C, WANG X, et al. Localization of passive intermodulation based on the concept of K-space multicarrier signal [J]. IEEE transactions on microwave theory and techniques, 2017, 65(12): 4997 – 5008. DOI: 10.1109/TMTT.2017.2705099
[11] LIU X H. Design of pulsed semiconductor laser ranging system [D]. Hohhot: Inner Mongolia University, 2014
[12] SHITVOV P A, ZELENCHUK E D, SCHUCHINSKY G A, et al. Passive intermodulation generation on printed lines: near-field probing and observations [J]. IEEE transactions on microwave theory and techniques, 2008, 56(12): 3121 – 3128, DOI: 10.1109/TMTT.2008.2007136

Biographies

DONG Anhua received his BS degree from Jilin University, China in 2021. He is currently pursuing his ME degree at University of Electronic Science and Technology of China. His research interests include the positioning of radiated passive inter-modulation.

LIANG Haodong received his PhD degree in guidance, navigation and control from University of Electronic Science and Technology of China in 2021. He is currently an algorithm engineer at ZTE Corporation. His research interests include the modeling and positioning of passive inter-modulation.

ZHU Shaohao received his PhD degree in underwater acoustic engineering from Northwestern Polytechnical University, China in 2021. He is currently an RF algorithm engineer at ZTE Corporation. His research interests include PIM mechanism, PIM cancellation, source location, and array beamforming.

ZHANG Qi received his BS degree from Henan Polytechnic University, China in 2023. He is currently pursuing his ME degree at University of Electronic Science and Technology of China. His research interest focuses on the positioning of radiated passive intermodulation.

ZHAO Deshuang (dszhao@uestc.edu.cn) received his PhD degree in optical engineering from University of Electronic Science and Technology of China (UESTC), in 2005. He is currently a professor with the School of Physics, UESTC. His research interests include the modeling and positioning of conductor-guided and radiated passive intermodulation.

Measurement and Analysis of Radar-Cross-Section of UAV at 21–26 GHz Frequency Band

AN Hao[1], LIU Ting[1], HE Danping[1], MA Yihua[2], DOU Jianwu[2]

(1. State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University, Beijing 100044, China;
2. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202501014

https://kns.cnki.net/kcms/detail/34.1294.TN.20250210.1609.002.html, published online February 11, 2025

Manuscript received: 2023-10-11

Abstract: With the emergence of the 6G technology, integrated sensing and communication (ISAC) has become a hot-spot vertical application. The low-altitude scenario is considered to be a significant use case of the ISAC. However, the existing channel model is hard to meet the demands of the sensing function. The radar-cross-section (RCS) is a critical feature for the sensing part, while accurate RCS data for the typical frequency band of ISAC are still lacking. Therefore, this paper conducts measurements and analysis of the RCS data of the unmanned aerial vehicles (UAVs) under multiple poses and angles in real flying conditions. The echo from a UAV is acquired in an anechoic chamber, and the RCS values are calculated. The results of different flying attitudes are analyzed, providing RCS features for the ISAC applications.

Keywords: unmanned aerial vehicle; radar-cross-section; integrated sensing and communication; anechoic chamber measurement

Citation (Format 1): AN H, LIU T, HE D P, et al. Measurement and analysis of radar-cross-section of UAV at 21 – 26 GHz frequency band [J]. ZTE Communications, 2025, 23(1): 107 – 114. DOI: 10.12142/ZTECOM.202501014

Citation (Format 2): H. An, T. Liu, D. P. He, et al., "Measurement and analysis of radar-cross-section of UAV at 21 – 26 GHz frequency band," *ZTE Communications*, vol. 23, no. 1, pp. 107 – 114, Mar. 2025. doi: 10.12142/ZTECOM.202501014.

1 Introduction

With the development of 6G, the integrated sensing and communication (ISAC) process is accelerating. An essential scenario for ISAC is the low-altitude environment, which includes specific use cases such as trunking communication and ad-hoc networks. Research on channel modeling for ISAC is now under lively discussion. However, the topic is mainly on the intelligent connected vehicles[1], and channel modeling for the low-altitude scenario with unmanned aerial vehicles (UAVs) as protagonists still needs to be further explored. The communications, sensing, and computing resources will be deeply integrated and mutually beneficial, providing efficient services for new intelligent applications such as intelligent transportation, UAV networks, space-air-ground-sea integrated networks, environmental detection, and metaverse[2 – 3]. The application potential of the low-altitude scenario is significant, as networks of UAVs can be used as sensor platforms to enable remote location coverage in emergencies like network impairment. Alternatively, UAVs can be used as low-cost infrastructure to provide traffic offload in crowded areas like stadiums[4]. Moreover, applying UAV networks to high-speed railways, especially in high-altitude unmanned areas, will effectively reduce maintenance costs. Due to their high mobility and low cost, UAVs play an increasingly significant role in many practical applications, including weather monitoring, forest fire detection, and emergency search and rescue[5]. By building a collaborative network architecture with multiple air base stations, it is possible to achieve multi-service, multi-access, multi-level coverage for post-disaster scenarios[6]. Besides, UAVs are expected to serve as an efficient complementary to terrestrial wireless communication systems to provide enhanced coverage and reliable connectivity to ground users[7]. UAVs facilitate more advanced technologies, such as federal learning[8], reconfigurable intelligent surface[9], and the Internet of Things[10]. Mobile edge computing (MEC) has developed into a promising computing paradigm. UAVs are practical in MEC since federated learning can improve the performance of UAV computing networks[11]. While there are ample application prospects and advantages, UAVs face significant challenges. Due to the high mobility, the requirements for dynamic channel modeling are demanding[12]. The channel status changes rapidly and should be updated frequently. Therefore, conducting the underlying theory and critical technology study for the UAV channels is of great importance and value.

This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. HC-CN-20220622006.

More research efforts have been put into current UAV channel modeling. Ref. [13] performs a channel replication of a 1 420 MHz air-to-air link using ray tracing and tapped-delay line models to describe the communication channel. The low detectability of radar targets is crucial for stealth. To avoid being detected by radar, the reduction technology of radar-cross-section (RCS) has become one of the research hotspots[14]. Currently, there is some literature on RCS measurement and analysis of UAVs. Most studies are mainly focused on the X-band (8 – 12 GHz) radars and even lower frequency bands. Ref. [15] discusses the relationship between RCS and the UAV flight range through dynamic measurements by a radar demonstrator system at 8.75 GHz. Ref. [16] presents the results of the measurement and analysis of several UAV RCS in different planes and from different elevation angles at 9 GHz. Ref. [17] describes the measurement and modeling of the dynamic RCS at 8 – 10 GHz and compares the difference between the probability of detection using dynamic and static RCSes. Due to the development of wireless communication frequency bands towards higher frequencies such as millimeter-wave, terahertz, and visible light bands, there will be more and more overlap with traditional sensing frequency bands. Besides, there is a lack of data support for sensory characteristics in the typical ISAC frequency band. RCS-based measurements at 15 GHz and 25 GHz are used for UAV recognition and detection in Refs. [18] and [19]. Diverse UAV detection and classification methods based on the RCS signatures are analyzed at 26 – 40 GHz[20 – 21]. However, accurate data on the UAVs for the ISAC frequency band is far from enough. It is worth noting that the structure and materials of the UAVs are different. Thus, the RCS values cannot be represented by a unified model. In addition, considering the variable attitude of the UAVs during flight and the unfixed relative positions between the UAVs and the base stations, conducting RCS measurements in practical scenarios is also challenging. Therefore, conducting multiple measurements and analyzing RCS values for different propeller states and attitudes while the UAV is stationary may be the most feasible approach. Channel models for communication functions can provide large- and small-scale channel parameters, while the contribution to sensing is weak. The sensing feature needs to be modeled accurately since the sensing function is considered a fundamental function of 6G networks. Collaborative perception can be achieved by deploying multiple UAVs[22]. From the existing literature, changes in the UAV structures have not attracted much attention. During the flight, the propeller states and attitudes are variable, which is also crucial for sensing. Therefore, accurate RCS data for the UAV structure changes at the key frequency band for communication sensing, i.e., 21 – 26 GHz, are essential and need to be further supplemented.

In response to the abovementioned demands and challenges, measurements are carried out for a typical UAV in this paper. Multi-angle bistatic measurements are conducted at different flight attitudes of 21 – 26 GHz. Accurate multi-angle RCSs are obtained after calibration and statistical analyses were performed. This study provides a data basis for the UAV application of ISAC and complements the missing measurement data of multiple attitudes and angles of UAVs in this frequency range. Besides, this work will contribute to the development of accurate ray-tracing simulation models for UAV scenarios[23 – 24], providing essential data support for ISAC channel standardization. The main contributions and novelties of this paper are as follows.

• An RCS measurement system is built based on a vector network analyzer (VNA) and a rotary table, which can measure the RCS and maximum received power of the UAV at any angle;

• The measurement data of RCS for quadcopter drones in the 21 – 26 GHz frequency band are filled;

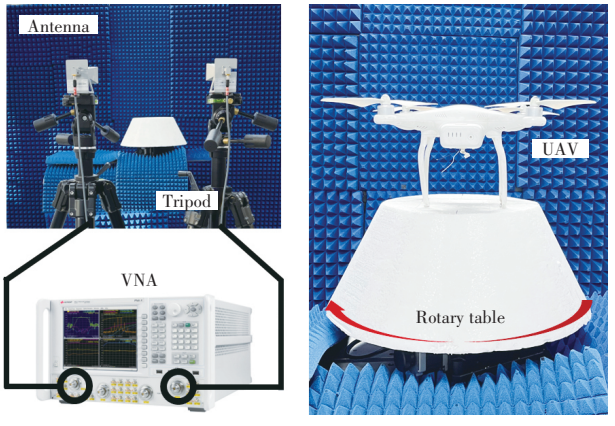
• The RCS of the UAV in different flight attitudes, propeller states, and angles between the transmitting and receiving antennas are measured, and the effects are compared.

The rest of this paper is organized as follows. Section 2 describes the RCS measurement system and layouts. The measurement results are introduced in Section 3, including the reference data, maximum received power, and RCS. Conclusions are drawn in Section 4.

2 Measurement Campaign

2.1 Measurement System

The RCS measurement of a UAV is carried out in an anechoic chamber to reduce the interference of external electromagnetic wave signals. At the same time, the absorbing materials reduce the multipath effects caused by the reflection of walls and ceilings. The measurement system consists of a VNA, a rotary table, two identical directional antennas, two tripods, and a quadcopter UAV, as shown in Fig. 1. The rotary table is made of low-density foam material, and its influence on electromagnetic wave propagation can be ignored. The UAV is placed on the rotary table and rotates synchronously with it.



UAV: unmanned aerial vehicle VNA: vector network analyzer

Figure 1. Proposed measurement system

The measurements are conducted at the center frequency of 23.5 GHz with a 5 GHz bandwidth. The frequency sampling number is 201, which indicates a frequency sampling resolution of 25 MHz. The gain of both directional antennas is 24.7 dBi to enhance the received signal strength. During the measurement process, the transceiver antenna is fixed on a tripod, and the UAV is placed on the rotary table to maintain the same antenna height. The heights of the Tx and the Rx from the ground are both 1.3 m, as well as the height of the UAV. The antennas and the UAV are at the same level to ensure that the antenna beams can cover the UAV. The rotary table is rotated at an interval of 5° during the measurements in order to measure the RCS at different angles as much as possible. The angles between the Tx and Rx are considered to be 10° and 45°. This is for comparing the RCS differences under different angles. More detailed measurement parameters are listed in Table 1.

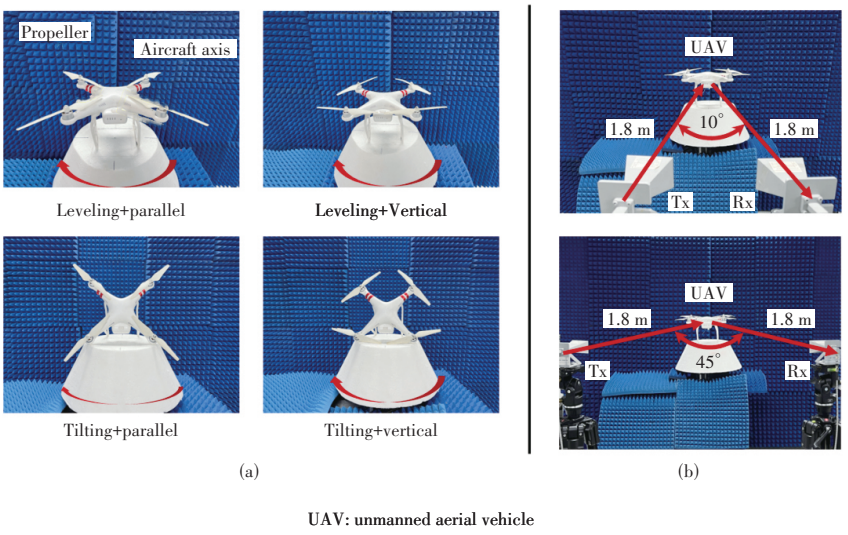


Figure 2. Layout of the antennas and the UAV:(a) attitudes of the UAV and (b) the positional relationship between the antennas and the UAV

2.2 UAV and Antenna Layouts

The UAV in the measurement is Phantom 3 Standard, which is one of the common camera drones. The diagonal size (propellers excluded) is 0.35 m. The length and width of the UAV are 0.25 m, and the height is 0.19 m.

Due to the difficulty of hovering the UAV, static measurements are carried out. To obtain the reflection and scattering characteristics of the various attitudes of the UAV, the fuselage states of leveling and tilting are considered. Besides, the propeller states of vertical and parallel to the aircraft axis are also considered. Thus, there are four types of UAV attitudes in the measurements in total, which are shown in Fig. 2a. Fig. 2b shows the positional relationship between the antennas and the UAV. All cases are summarized in Table 2. The distance between Tx and the UAV is 1.8 m. The distance between the UAV and Rx is 1.8 m. The diameters of the aircraft axes are 0.03 m. The measurement meets the far-field conditions, according to the following equation:

$$d_f = 2D^2/\lambda \quad (1),$$

where d_f is the distance of the Fraunhofer region, D is the maximum linear dimension of the antenna, and λ is the wavelength.

Table 1. Measurement configuration

Measurement Parameters	Values
Center frequency	23.5 GHz
Bandwidth	5.0 GHz
Frequency samples	201.0
Rotation angle interval	5.0°
Tx and Rx heights from the ground	1.3 m
UAV height from the ground	1.3 m
Antenna gain	24.7 dBi
Angle between Tx and Rx	10.0°/45.0°

UAV: unmanned aerial vehicle

Table 2. Measurement cases

Case	UAV States	Propeller States to Aircraft Axis	Angle Between Tx and Rx
1	Leveling	Parallel	10°
2	Leveling	Vertical	
3	Tilting	Parallel	
4	Tilting	Vertical	
5	Leveling	Parallel	45°
6	Leveling	Vertical	
7	Tilting	Parallel	
8	Tilting	Vertical	

UAV: unmanned aerial vehicle

3 Measurement Results

3.1 Reference Data

To obtain accurate antenna gains, measurements are conducted under line of sight (LoS) conditions. Besides, the case without placing the UAV is measured to provide a reference. The distance between the Tx and Rx is 3.6 m, which is twice the distance from the Tx or Rx to the UAV in the RCS measurements. Moreover, the power delay profiles (PDPs) with and without the UAV are compared, which is shown in Fig. 3.

Fig. 3 shows that the delays corresponding to the strongest power are the same in all cases. The power is the highest at

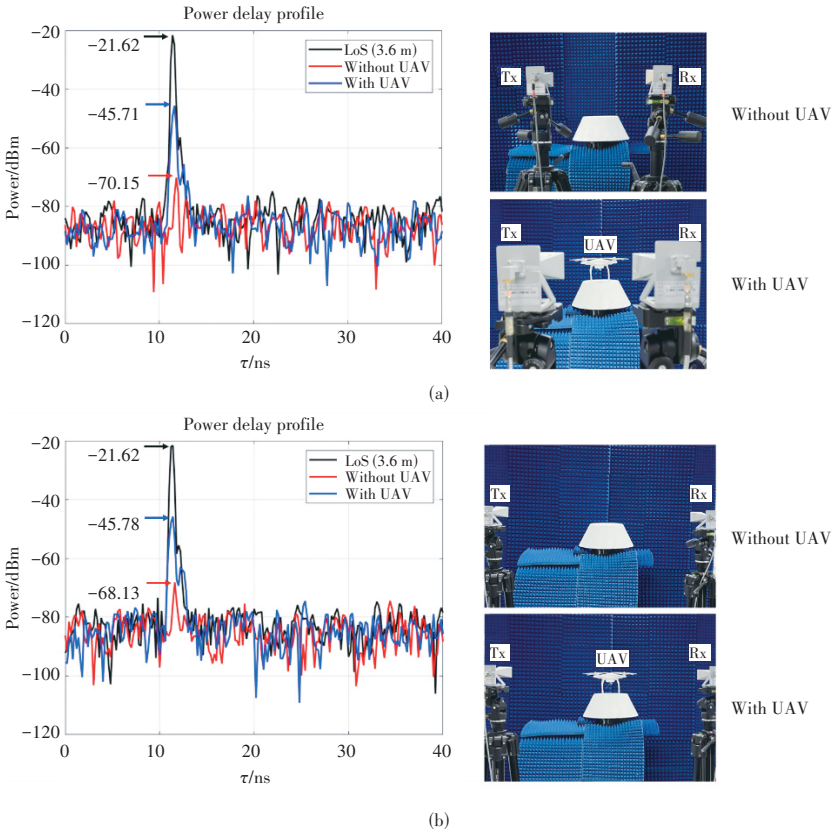


Figure 3. PDP comparison of the LoS and w/o the UAV at the Tx-Rx angle of (a) 10° and (b) 45°

the LoS and the lowest when there is no UAV placed, which indicates that the absorbing materials are very effective. Since Tx and Rx are the same, the antenna gain G_{Ant} can be calculated by:

$$G_{\mathrm{Ant}} = (P_{\max} + \mathrm{FSPL}_{23.5\,\mathrm{GHz},\,3.6\,\mathrm{m}})/2 \quad (2),$$

where $P_{\max}$ denotes the maximum power of the PDP and $\mathrm{FSPL}_{23.5\,\mathrm{GHz},\,3.6\,\mathrm{m}}$ denotes the free space path loss at 3.6 m at the frequency of 23.5 GHz.

3.2 Maximum Received Power

The maximum power at each measurement angle is analyzed before RCS, which can help us gain a preliminary understanding of the reflection characteristics of the UAV at different angles. For each attitude of the UAV at two angles, the measurements are conducted every 5° of the rotation of the rotary table, and thus 72 measurement results are obtained. The values of the maximum power are found from all measured angles. The radar charts are shown in Fig. 4, where the maximum power is higher at the four measured angles of 0°, 90°, 180°, and 270°, which is mainly caused by the reflection of the battery module below the UAV. The power at the same measurement angle varies slightly at different attitudes. In addition, the maximum power of Tx and Rx at the angle of 10° is generally higher than that at the angle of 45°.

3.3 RCS

The radar acquires the target information by processing the echo data. Therefore, the design and operation of radars are critical to quantify and describe the echo, especially in terms of target characteristics such as the size, shape, and orientation. For that purpose, the target is ascribed to an effective area called the RCS[25]. The RCS of the target is the ratio of the power scattered back to the radar receiver over the incident radar power density per unit of solid angles on the target, which is expressed as follows[26]:

$$\sigma = \frac{P_r (4\pi)^3 d_t^2 d_r^2}{P_t G_t G_r \lambda^2} \quad (3),$$

where σ represents the RCS, P_t and P_r represent the transmitted power and received power, d_t and d_r represent the distance from the target to Tx and Rx, and G_t and G_r represent the gain of Tx and Rx. Fig. 5 shows the RCS results and cumulative distribution function (CDF) at each measured angle in the cases of two different angles between Tx and Rx. All the results are summarized in Table 3.

1) Case 1

When the angle between the Tx and Rx is 10°, the UAV is leveling, and the propellers are parallel to the aircraft axis. The maximum value of RCS is obtained at 0° of the rotation in this condition, where the UAV faces the antennas. The battery module below the UAV generates a strong echo. The CDF shows that the mean value of the RCS in this case is about −31.68 dBsm.

2) Case 2

When propellers are vertical to the aircraft axis, the maximum value of RCS is also obtained at 0° of the rotation. Compared with the previous UAV attitude scenario, values of the RCS are slightly different from those at the measured angles of 0°, 90°, 180°, and 270°. However, there are significant differences in other measured angles. Because, at those angles, the propellers are in the lobes of the Tx and Rx, significantly impacting the echo. From the CDF, the mean value of RCS in this case is also about −31.63 dBsm.

3) Case 3

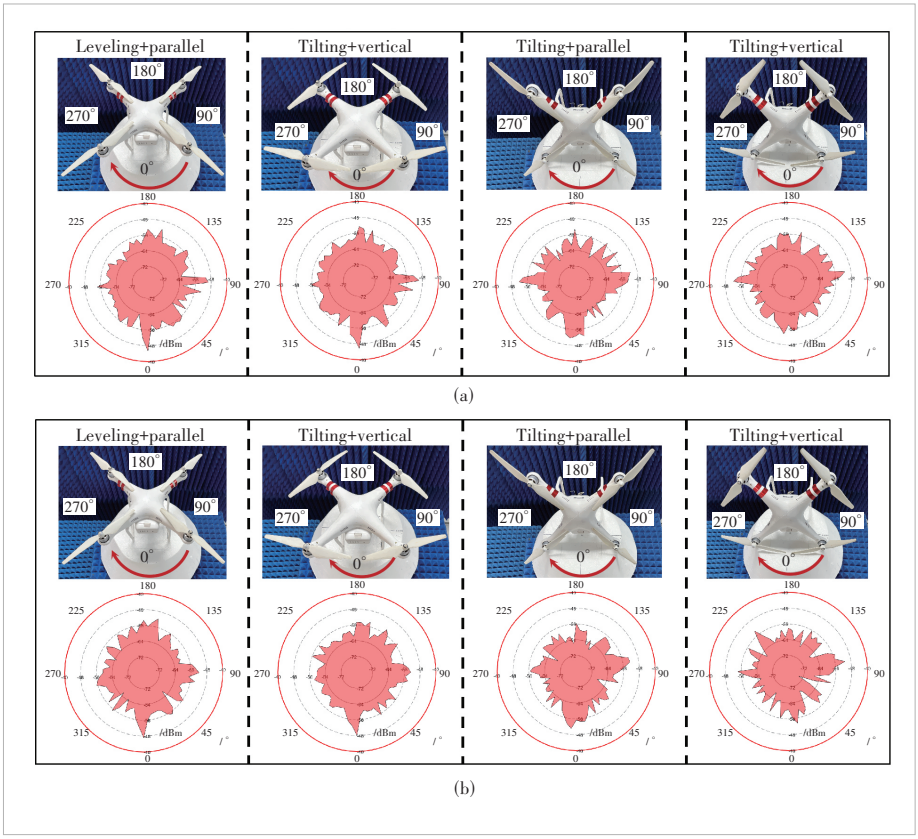


Figure 4. Maximum power at different attitudes at the Tx-Rx angles of (a) 10° and (b) 45°

When the angle between the Tx and Rx is still 10°, the UAV changes to tilt, and the propellers are parallel to the aircraft axis. Like the previous ones, the maximum value of the RCS is also obtained at 0°, indicating that the reflecting surface on the top of the UAV still plays a crucial role. However, this value significantly decreases as the size of the reflection surface in this case is smaller than that of the leveling ones. Besides, the values of the RCS fluctuate more sharply at other measured angles because of the UAV structure. Significant differences in the structure of the UAV at different angles lead to rapid changes in the size and shape of the reflecting surfaces. The mean value of the RCS in this case is about −33.27 dBsm.

Table 3. RCS measurement results

Case	RCS		
	Maximum Value/dBsm (Corresponding Angle)	Minimum Value/dBsm (Corresponding Angle)	Mean Value/ dBsm
1	−19.01 (0°)	−38.81 (225°)	−31.68
2	−19.40 (0°)	−36.53 (45°)	−31.63
3	−25.12 (350°)	−40.87 (215°)	−33.27
4	−24.72 (100°)	−39.50 (290°)	−32.85
5	−20.78 (0°)	−40.33 (45°)	−32.09
6	−21.78 (0°)	−36.95 (255°)	−32.36
7	−24.71 (355°)	−42.93 (240°)	−34.02
8	−24.59 (100°)	−47.95 (65°)	−34.63

RCS: radar-cross-section

4) Case 4

As for the case of the propellers being vertical to the aircraft axis, the maximum value of RCS is obtained at the measured angle of 90°. The rotation of the propellers affects the size of the reflecting surface. Compared with the case where the UAV is leveling and the propellers are vertical, the value of the RCS significantly decreases at 0°. Meanwhile, the values fluctuate more sharply at other measurement angles for the same reason as the previous one. The mean value of the RCS is about −32.85 dBsm according to the CDF.

5) Case 5

Fig. 5b shows the results of the RCS when the angle between the Tx and Rx is 45°. The maximum value of RCS is obtained at 0° while the UAV is leveling and the propellers are parallel to the aircraft axis. The RCS values at most measured angles slightly increase compared with those at 10°, given the same attitude, indicating that the reflection area is larger at this angle. The mean value of the RCS of the UAV in this case is about −32.09 dBsm.

6) Case 6

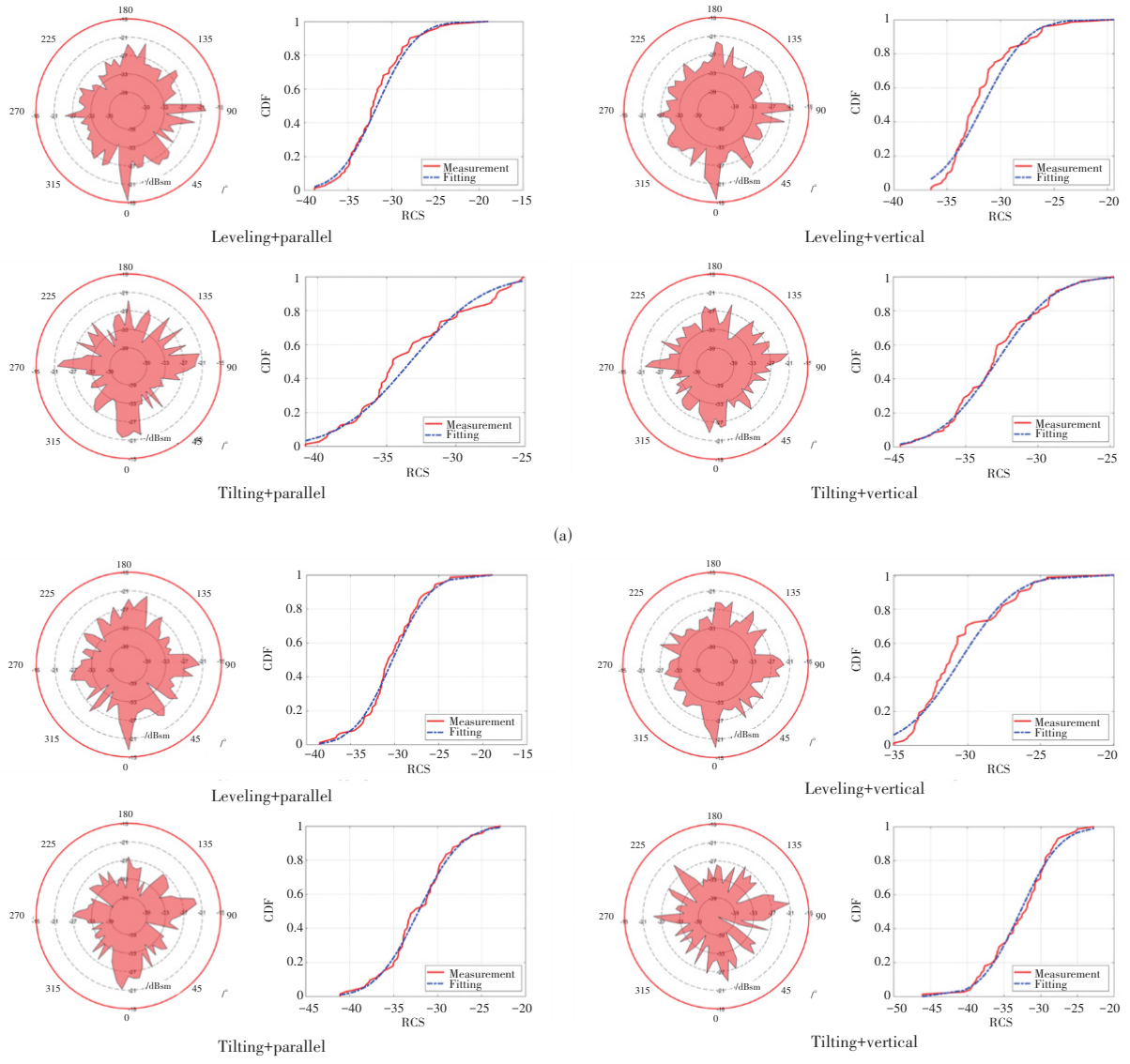
A similar situation occurs when the UAV is tilting and the propellers are vertical to the aircraft axis. The maximum value of the RCS is also obtained at 0°. Moreover, the mean value of the RCS is −32.36 dBsm.

7) Case 7

When the angle between the Tx and Rx is 45°, the UAV tilts and the propellers are parallel to the aircraft axis, and the maximum value of the RCS is also obtained at 0°. Besides, the values of the RCS show little change in most measured angles. The values of the RCS are mainly smaller than the case of the UAV leveling and the propellers are parallel to the aircraft axis in most measured angles because of the significant impact on the reflection surface. According to the CDF, the mean value of the RCS in this case is about −34.02 dBsm.

8) Case 8

As the UAV is tilting and the propellers are vertical to



CDF: cumulative distribution function RCS: radar-cross-section

Figure 5. RCS of the UAV at different attitudes at the Tx-Rx angles of (a) 10° and (b) 45°

the aircraft axis, the maximum value of RCS is obtained at 90°, same with the situation when the angle between the Tx and Rx is 10°. However, the values of RCS show little difference from the situation where the angle between the Tx and Rx is 10° in most measured angles. Compared with the case of the UAV leveling and the propellers being vertical to the aircraft axis at the same angle between the Tx and Rx, most values of the RCS decrease as the tilt of the UAV at this angle causes a more significant impact on the reflection surface. Furthermore, the mean value of the RCS of the UAV is about −34.63 dBsm in this case.

4 Conclusions

In this paper, the UAV RCS measurements are conducted based on the VNA. The UAV is measured in all directions by using the rotary table. The angles between the Tx and Rx include 10° and 45°. Four types of UAV attitudes are considered for a comprehensive analysis. The UAV is leveling and tilting, and the propellers are parallel and vertical to the aircraft axis.

As for the measurement results, the maximum received power and the RCS are analyzed. For complex targets like UAVs, there is no fixed calculation relationship between their fuselage structure and RCS. It is found that the maximum

power is mainly affected by the reflection of the main part of the UAV rather than the attitudes. The power at the same measurement angle varies slightly at different attitudes. The angle between the Tx and Rx can also influence the results, with a smaller angle resulting in a higher maximum power. The maximum value of the RCS is mostly measured at around 0°, except for the case where the UAV is tilting and the propellers are vertical to the aircraft axis, which is obtained at 100°. The mean values of the RCS in different cases are between −31 dBsm and −35 dBsm. As the angle between the Tx and Rx increases, the values of RCS generally decrease. In addition, the attitude change will significantly impact the changes in the values of the RCS at different measured angles.

This paper provides reference data at the millimeter wave band for studying the ISAC channel of UAVs. The flight status of UAVs may be determined by constantly detecting RCS values. In addition, the results of this paper can also be used as data references for the ray-tracing simulation of UAVs[27]. Using the same method, more comprehensive measurements of multiple types of unmanned aerial vehicles, multiple frequency bands, and multiple flight attitudes can be conducted. The RCS models of different UAVs at different attitudes and incident angles can be explored in the future.

References

[1] CHENG X, DUAN D L, GAO S J, et al. Integrated sensing and communications (ISAC) for vehicular communication networks (VCN) [J]. IEEE Internet of Things journal, 2022, 9(23): 23441 – 23451. DOI: 10.1109/JIOT.2022.3191386
[2] WANG C X, YOU X H, GAO X Q, et al. On the road to 6G: visions, requirements, key technologies, and testbeds [J]. IEEE communications surveys & tutorials, 2023, 25(2): 905 – 974. DOI: 10.1109/COMST.2023.3249835
[3] WU H C, LI H J, TAO X F. Green air-ground integrated heterogeneous network in 6G era [J]. ZTE communications, 2021, 19(1): 39 – 47. DOI: 10.12142/ZTECOM.202101006
[4] ALZAHRANI B, OUBBATI O S, BARNAWI A, et al. UAV assistance paradigm: state-of-the-art in applications and challenges [J]. Journal of network and computer applications, 2020, 166: 102706. DOI: 10.1016/j.jnca.2020.102706
[5] ZENG Y, ZHANG R, LIM T J. Wireless communications with unmanned aerial vehicles: opportunities and challenges [J]. IEEE communications magazine, 2016, 54(5): 36 – 42. DOI: 10.1109/MCOM.2016.7470933
[6] HE Y X, WANG D W, HUANG F H, et al. A V2I and V2V collaboration framework to support emergency communications in ABS-aided Internet of vehicles [J]. IEEE transactions on green communications and networking, 2023, 7(4): 2038 – 2051. DOI: 10.1109/TGCN.2023.3245098
[7] LIU T X, MIN S, LYU R L, et al. UAV assisted heterogeneous wireless networks: potentials and challenges [J]. ZTE communications, 2018, 16(2): 3 – 8. DOI: 10.3969/J.ISSN.1673-5188.2018.02.002
[8] WANG P F, SONG W, SUN G, et al. Air-ground integrated low-energy federated learning for secure 6G communications [J]. ZTE Communications, 2022, 20(4): 32 – 40. DOI: 10.12142/ZTECOM.202204005
[9] SHEN Y, OU P, CHEN F K, et al. Reconfigurable intelligent surface-assisted channel characteristics in 5G high-speed railway scenario [J]. Journal of Beijing Jiaotong University, 2023, 47(2): 23 – 35. DOI: 10.11860/j.issn.1673-0291.20220098
[10] XU S Y, LYU J S. Maximizing UAV coverage efficiency based on retransmission in the Internet of Things [J]. Journal of Beijing Jiaotong University, 2023, 47(2):58 – 66. DOI: 10.11860/j.issn.1673-0291.20220133
[11] ALSAMHI S H, SHVETSOV A V, KUMAR S, et al. Computing in the sky: a survey on intelligent ubiquitous computing for UAV-assisted 6G networks and industry 4.0/5.0 [J]. Drones, 2022, 6(7): 177. DOI: 10.3390/drones6070177
[12] CHENG X, HUANG Z W, BAI L. Channel nonstationarity and consistency for beyond 5G and 6G: a survey [J]. IEEE communications surveys & tutorials, 2022, 24(3): 1634 – 1669. DOI: 10.1109/COMST.2022.3184049
[13] AN H, GUAN K, LI W B, et al. Measurement and ray-tracing for UAV air-to-air channel modeling [C]//Proceedings of IEEE 5th International Conference on Electronic Information and Communication Technology (ICEICT). IEEE, 2022: 415 – 420. DOI: 10.1109/ICEICT55736.2022.9908966
[14] WANG W J, SHI Y, MENG Z K, et al. A metasurface design method for dual wide band radar cross section reduction [J]. Chinese journal of radio science, 2021, 36(6): 887 – 895
[15] DE QUEVEDO Á D, URZAIZ F I, MENOYO J G, et al. Drone detection with X-band ubiquitous radar [C]//Proceedings of 19th International Radar Symposium (IRS). IEEE, 2018: 1 – 10. DOI: 10.23919/IRS.2018.8447942
[16] SEDIVY P, NEMEC O. Drone RCS statistical behaviour [EB/OL]. (2021-11-07) [2023-04-16]. https://www.sto.nato.int/publications/STO%20Meeting%20Proceedings/STO-MP-MSG-SET-183/MP-MSG-SET-183-04.pdf
[17] GUAY R, DROLET G, BRAY R J. Measurement and modelling of the dynamic radar cross-section of an unmanned aerial vehicle [J]. IET radar, sonar & navigation, 2017, 11(7): 1155 – 1160
[18] EZUMA M, ANJINAPPA C K, SEMKIN V, et al. Comparative analysis of radar-cross-section- based UAV recognition techniques [J]. IEEE sensors journal, 2022, 22(18): 17932 – 17949. DOI: 10.1109/JSEN.2022.3194527
[19] EZUMA M, ANJINAPPA C K, FUNDERBURK M, et al. Radar cross section based statistical recognition of UAVs at microwave frequencies [J]. IEEE transactions on aerospace and electronic systems, 2021, 58(1): 27 – 46. DOI: 10.1109/TAES.2021.3096875
[20] SEMKIN V, YIN M S, HU Y Q, et al. Drone detection and classification based on radar cross section signatures [C]//International Symposium on Antennas and Propagation (ISAP). IEEE, 2021: 223 – 224
[21] SEMKIN V, HAARLA J, PAIRON T, et al. Analyzing radar cross section signatures of diverse drone models at mmWave frequencies [J]. IEEE access, 2020, 8: 48958 – 48969. DOI: 10.1109/ACCESS.2020.2979339
[22] LIAO N W, QIAN P Z, CHEN Y, et al. A joint planning method for the number of UAVs and spectrum resource in perceptual missions [J]. Chinese journal of radio science, 2023, 38(5):764 – 772. DOI: 10.12265/j.cjors.2022212
[23] HE D P, AI B, GUAN K, et al. The design and applications of high-performance ray-tracing simulation platform for 5G and beyond wireless communications: a tutorial [J]. IEEE communications surveys & tutorials, 2018, 21(1): 10 – 27. DOI: 10.1109/COMST.2018.2865724.
[24] LIN Y C, ZHONG Z D, GUAN K, et al. Channel characteristic of millimeter wave massive MIMO under train-to-infrastructure scenario based on ray-tracing method [J]. Chinese journal of radio science, 2017, 32(5): 595 – 601. DOI: 10.13443/j.cjors.2017080902
[25] KNOTT E F, SCHAEFFER J F, TULLEY M T. Radar cross section [M]. Henderson, USA: SciTech Publishing, 2004
[26] HE D P, GUAN K, AI B, et al. Channel measurement and ray-tracing simulation for 77 GHz automotive radar [J]. IEEE transactions on intelligent transportation systems, 2022, 24(7): 7746 – 7756. DOI: 10.1109/TITS.2022.3208008
[27] HE D P, GUAN K, YAN D, et al. Physics and AI-based digital twin of

multi-spectrum propagation characteristics for communication and sensing in 6G and beyond [J]. IEEE journal on selected areas in communications, 2023, 41(11): 3461 – 3473. DOI: 10.1109/JSAC.2023.3310108

Biographies

AN Hao received his BE degree in communication engineering from Beijing Information and Science Technology University, China in 2021. He is currently pursuing his PhD degree with the School of Electronic and Information Engineering, Beijing Jiaotong University, China. He received the Best Paper Award in ICEICT 2022. His current research interests include measurement and modeling of wireless propagation channels, UAV communications, and integrated sensing and communications.

LIU Ting received her BE degrees from both Beijing Jiaotong University, China and Lancaster University (with first-class honors), UK in 2021, in electronic and communication engineering. She is currently working toward her doctoral degree with the School of Electronic and Information Engineering, Beijing Jiaotong University. Her current research interests include radio propagation, deterministic channel modelling, and hybrid channel modelling for ISAC.

HE Danping (hedanping@bjtu.edu.cn) received her BE degree from Huazhong University of Science and Technology, China in 2008, MS degree from the Universite Catholique de Louvain, Belgium and Politecnico di Torino, Italy in 2010, and PhD degree from Universidad Politécnica de Madrid, Spain in 2014. In 2012, she was a visiting scholar with Institut National de Recherche en Informatique et en Automatique, France. From 2014 to 2015, she was a research engineer with Huawei Technologies. From 2016 to 2018, she was a postdoctoral researcher with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, China. She is currently with Beijing Jiaotong University as an associate professor. She has authored or coauthored more than 60 papers, three patents and one IEEE standard. Her research interests include radio propagation and channel modeling, ray tracing simulator development and wireless communication algorithm design. She is a member of the CA15104 Initiative, and was the recipient of five Best Paper Awards.

MA Yihua received his BE degree from Southeast University, China in 2015 and MS degree from Peking University, China in 2018. He has been with ZTE Corporation since 2018, where he is now an expert-level research engineer in the Department of Algorithm and a member of the State Key Laboratory of Mobile Network and Mobile Multimedia Technology, China. His main research interests include mMTC, joint communication and sensing, grant-free transmission, and massive MIMO.

DOU Jianwu received his PhD degree in robotic mechanism from Beijing University of Technology, China in 2001. From 2000 to 2014, he was the leader of the Wireless RRM Team, ZTE Corporation including 2G/3G/4G/WLAN and was in charge of developing the multi-RAT wireless system simulation platform. From 2012 to 2014, he was the product manager of ZTE iNES, a multi-cell/multi-UE hardware wireless channel emulator. From 2005 to 2017, he was the vice director of Wireless Algorithm Department of ZTE Corporation. His current research interests are B5G/6G channel modeling, new air-interface, unmanned aerial vehicles, non-terrestrial network research, THz, meta-materials and RIS. Dr. DOU received the Science and Technology Award (the 1st Level) in 2014/2015 and the Award for Chinese Outstanding Patented Invention in 2011 from China Institute of Communications and WIPO-SIPO, respectively.

Doppler Rate Estimation for OTFS via Large-Scale Antenna Array

SHAN Yaru[1], WANG Fanggang[1], HAO Yaxing[1], HUA Jian[2], XIN Yu[2]

(1. Beijing Jiaotong University, Beijing 100044, China;
2. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202501015

https://kns.cnki.net/kcms/detail/34.1294.TN.20250103.1627.002.html, published online January 6, 2025

Manuscript received: 2024-03-22

Abstract: Orthogonal time frequency space (OTFS) can resist the Doppler effect and guarantee reliable communication in high-speed scenarios. However, the Doppler rate induced by the relative acceleration between the transmitter and receiver degrades the performance of the OTFS. So far, the impact of the Doppler rate on OTFS systems has not been addressed. In this paper, we first introduce the Doppler rate in the OTFS system and derive the delay-Doppler domain input-output relation. In addition, the impact of the Doppler rate on the effective delay-Doppler domain channel is characterized by utilizing the first mean value theorem for definite integrals to avoid complicated integrals. To mitigate the effect of the Doppler rate, a large-scale antenna array is arranged at the receiver to separate each path of the multi-path channel through a high-resolution spatial matched filter beamformer. Next, the Doppler rate estimation scheme for an arbitrary order Doppler rate is proposed based on the successive interference cancellation pattern and the maximization of the spectrum of the ratio of high-order moments between the received samples in the identified branch and the transmitted samples. Finally, the estimation accuracy of the Doppler rate and the error performance of the proposed transceiver are validated by the numerical results.

Keywords: beamforming; Doppler rate; OTFS

Citation (Format 1): SHAN Y R, WANG F G, HAO Y X, et al. Doppler rate estimation for OTFS via large-scale antenna array [J]. ZTE Communications, 2025, 23(1): 115 - 122. DOI: 10.12142/ZTECOM.202501015

Citation (Format 2): Y. R. Shan, F. G. Wang, Y. X. Hao, et al., "Doppler rate estimation for OTFS via large-scale antenna array," *ZTE Communications*, vol. 23, no. 1, pp. 115 - 122, Mar. 2025. doi: 10.12142/ZTECOM.202501015.

1 Introduction

Orthogonal time frequency space (OTFS) modulation has been proposed to overcome the high Doppler effect[1-2]. Specifically, the information-bearing symbols are modulated in the delay-Doppler domain rather than the time-frequency domain of the orthogonal frequency division multiplexing. Each symbol in the delay-Doppler domain is transformed into the whole time-frequency domain by the two-dimensional inverse symplectic finite Fourier transform (ISFFT), which enables the OTFS to harness the full diversity[3]. The delay-Doppler domain captures the physical characteristics including the delay shifts and the Doppler shifts of the channel, which enables the sparsity of the channel[4]. The maximum delay and the maximum Doppler shift are within the corresponding range of the delay-Doppler domain. Therefore, the delay-Doppler domain channel is underspread and quasi-stationary. The sparse and relatively quasi-stationary characteristics in the delay-Doppler domain benefit the channel estimation and the data detection tasks for OTFS.

In mobile radio transmission scenarios, such as the radar detection, low earth orbit (LEO) satellites, and millimeter-wave systems, the received signal may experience significant time-varying Doppler distortion due to the relative motion between the transceivers[5-8]. Then the Doppler rate as a high-order motion parameter related to the motion acceleration rate must be considered in the system model. Thus, the Doppler shift is no longer a constant but a variable that changes with time. Various methods are adopted to address the Doppler rate under different scenarios. To image a ground-moving target with a synthetic radar system, the third-order Doppler frequency mitigation schemes are designed. The coherent integration detection schemes based on the Keystone transform and even the second-order Keystone transform, and the generalized Hough-high-order ambiguity function are proposed in Refs. [5] and [6], respectively. In Ref. [7], a new fast Doppler shift and Doppler rate joint acquisition method derived from the spectrum method is proposed for hypersonic vehicle communications. Based on the sequential importance sampling,

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant No. 2022JBQY004, the National Natural Science Foundation of China under Grant Nos. 62471026 and 62221001, the Zhongguancun XinXi Disruptive Technology Innovation Foundation under Grant No. ZZ-2024-001, the Joint Funds for Railway Fundamental Research of National Natural Science Foundation of China under Grant No. U2368201, the Postdoctoral Fellowship Program and China Postdoctoral Science Foundation BX20240471, and ZTE Industry-University-Institute Cooperation Funds under Grant No. HC-CN-03-2019-12.

the joint estimation of the carrier phase, Doppler shift, Doppler rate, and data detection using particle filters is proposed in Ref. [8]. However, the above-mentioned schemes only consider the Doppler rate in the scenario where only a single path exists and is not always valid in the common communication systems. Moreover, to the best of our knowledge, there does not exist any open literature that introduces and then addresses the Doppler rate effect for OTFS.

In this paper, we first introduce the Doppler rate in the OTFS system. The delay-Doppler domain input-output relation is derived under an arbitrary-order Doppler rate. In addition, the Doppler rate effect is characterized by taking advantage of the first mean value theorem for definite integrals. Then, the system model is proposed by arranging a large-scale antenna array at the receiver and a joint frame structure is considered, where the first transmission frame is utilized to estimate the Doppler rate and then the Doppler rate is compensated in the subsequent frames through precoding. Next, the Doppler rate estimation and compensation scheme that applies to the system with an arbitrary-order Doppler rate and performs in the first frame is proposed. Although some studies utilize the angle domain to separate scattering paths, the existing literature does not account for the Doppler rate effect[9]. The main contributions of this paper are summarized as follows:

1) The Doppler rate is first introduced in the OTFS system. Based on the introduction of the effect of the Doppler rate, the delay-Doppler domain input-output relation is rederived and the influence of the Doppler rate is characterized by utilizing the first mean value theorem for definite integrals.

2) The receiver scheme is designed with a large-scale antenna array to estimate the Doppler rate in each identified branch. Specifically, the different scattering paths are separated in the angle domain to create the signal path condition to simplify the Doppler rate estimation.

3) The Doppler rate mitigation scheme applied to the arbitrary order Doppler rate is proposed by utilizing the maximization of the spectrum of the ratio of high-order moments between the received samples in the identified branch and the transmitted samples.

The remainder of the paper is organized as follows. Section 2 introduces the system model. Section 3 introduces the OTFS transceiver, followed by the Doppler rate estimation in Section 4. The simulation results and conclusions are provided in Sections 5 and 6, respectively.

2 System Model

We first introduce the Doppler rate effect to the delay-Doppler domain channel. Then, the generalized delay-Doppler domain input-output relation for an OTFS system with the arbitrary-order Doppler rate is derived. Finally, the effect of the Doppler rate is characterized.

2.1 Channel with Doppler Rate*

The delay-Doppler domain channel with the Doppler rate is sparse and is expressed as:

$$h(\tau,\nu,t)=\sum_{p=0}^{P-1}\beta_p\delta(\tau-\tau_p)\delta(\nu-\nu_p-\sum_{q=1}^{Q}a_q t^q) \tag{1}$$

where P is the total number of the channel taps; β_p and τ_p are the channel coefficient and the delay shift of the p-th path, respectively; $\nu_p=f_d\cos\theta_p$ where f_d is the maximum Doppler shift and θ_p is the angle of arrival of the p-th path; Q is the highest order of the Doppler rate and a_q is the coefficient of the q-th order Doppler rate. The variation of the delay can exist in the high-speed scenarios as in Ref. [10].

2.2 Input-Output Relation in Delay-Doppler Domain

Without the noise, the delay-Doppler domain input-output relation can be expressed as:

$$y[k,l]=\frac{1}{MN}\sum_{k'=0}^{N-1}\sum_{l'=0}^{M-1}x[k',l']h_{k,l}[k',l'],\quad k\in\mathcal{I}_N,l\in\mathcal{I}_M \tag{2}$$

where $x[k',l']$ and $y[k,l]$ are the information-bearing symbols and the received symbols in the delay-Doppler domain, respectively; N and M are the numbers of the samples in the Doppler and the delay domain, respectively; $\mathcal{I}_N=[0,1,\cdots,N-1]$ is defined as shorthand hereafter to represent an index set; $h_{k,l}[k',l']$ is the sampled effective delay-Doppler domain channel and can be expressed as:

$$h_{k,l}[k',l']=\underbrace{\sum_{n=0}^{N-1}\sum_{m=0}^{M-1}\sum_{m'=0}^{M-1}H_{n,m}[n,m']e^{-j2\pi n\left(\frac{k-k'}{N}\right)}e^{j2\pi\frac{ml-m'l'}{M}}}_{\triangleq h_{k,l}^{(1)}[k',l']}+$$
$$\underbrace{\sum_{n=1}^{N-1}\sum_{m=0}^{M-1}\sum_{m'=0}^{M-1}H_{n,m}[n-1,m']e^{-j2\pi\left(\frac{nk}{N}-\frac{(n-1)k'}{N}\right)}e^{j2\pi\frac{ml-m'l'}{M}}}_{\triangleq h_{k,l}^{(2)}[k',l']} \tag{3}$$

We denote $H_{n,m}[n',m']$ as the sampled time-frequency domain effective channel. With the rectangular pulses considered for the transmit and the receive pulse-shaping functions, $H_{n,m}[n',m']$ is nonzero when $n'=n$ or $n'=n-1$, and $H_{n,m}[n,m']$ and $H_{n,m}[n-1,m']$ are expressed as:

$$H_{n,m}[n,m']=$$
$$\sum_{p=0}^{P-1}\frac{\beta_p}{T}\int_0^{T-\tau_p}e^{j2\pi\left(\nu_p+\sum_{q=1}^{Q}a_q(t''+\tau_p+nT)^q\right)(t''+nT)}\times$$
$$e^{j2\pi m'\Delta f t''}e^{-j2\pi m\Delta f(t''+\tau_p)}\mathrm{d}t'' \tag{4}$$

* We pay attention to the time-varying Doppler shift in this manuscript and the effect of the time-varying delay will be addressed in our future work.

$$H_{n,m}[n-1,m'] = \sum_{p=0}^{P-1}\frac{\beta_p}{T}\int_{T-\tau_p}^{T} e^{j2\pi\left(\nu_p+\sum_{q=1}^{Q}a_q(t''+\tau_p+(n-1)T)^q\right)(t''+(n-1)T)} \times e^{j2\pi m'\Delta f t''}e^{-j2\pi m\Delta f(t''+\tau_p)}\mathrm{d}t'' \quad (5).$$

We can see from Eqs. (3) – (5) that the calculation of the $h_{k,l}[k',l']$ involves the integral of the function with the form $\int_{a_1}^{b_1} e^{t''^{Q+1}}\mathrm{d}t''$, which is difficult to calculate directly.

2.3 Doppler Rate Effect Characterization

To demonstrate the influence of the Doppler rate, the first mean value theorem for definite integrals is utilized to avoid the complicated calculation in Eqs. (4) and (5). We assume that $\exists\xi_1 \in [0, T-\tau_p]$ and $\exists\xi_2 \in [T-\tau_p, T]$, and Eqs. (6) and (7) hold.

From Eq. (3), we can further express $h_{k,l}^{(1)}[k',l']$ and $h_{k,l}^{(2)}[k',l']$ as:

$$H_{n,m}[n,m'] = \sum_{p=0}^{P-1}\frac{\beta_p(T-\tau_p)}{T}e^{j2\pi m'\Delta f\xi_1}e^{-j2\pi m\Delta f(\xi_1+\tau_p)} \times e^{j2\pi\left(\nu_p+\sum_{q=1}^{Q}a_q(\xi_1+\tau_p+nT)^q\right)(\xi_1+nT)} \quad (6),$$

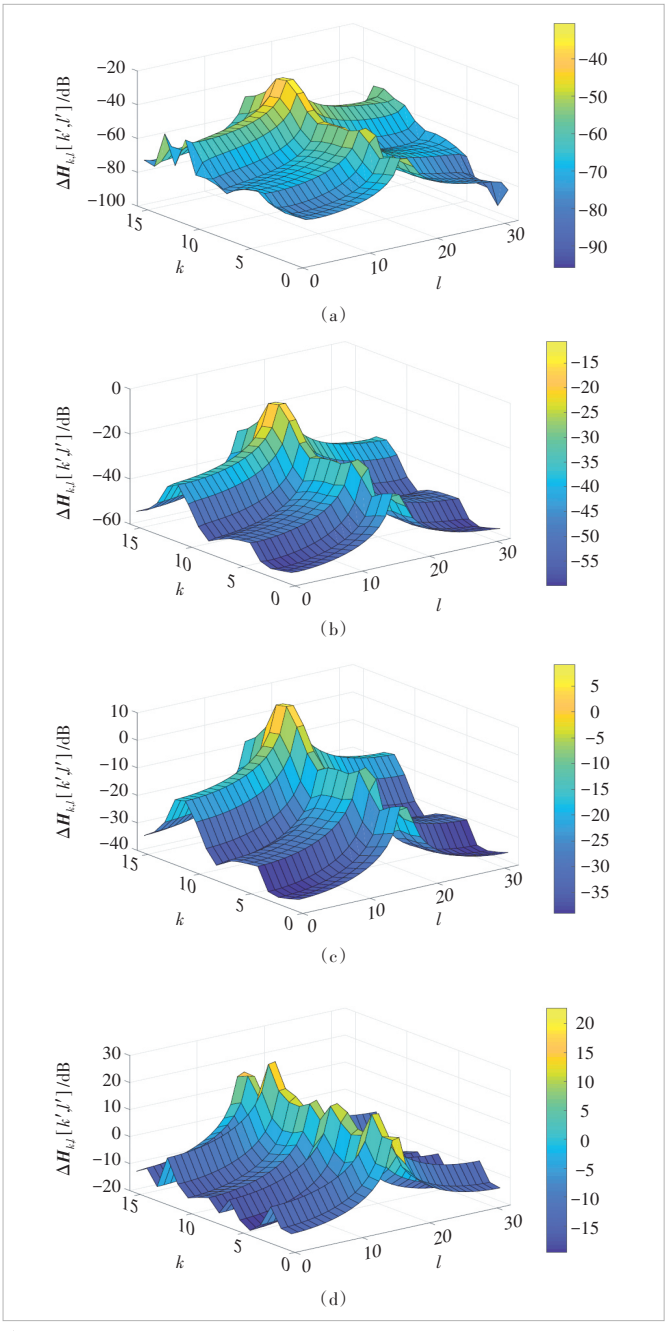
$$H_{n,m}[n-1,m'] = \sum_{p=0}^{P-1}\frac{\beta_p\tau_p}{T}e^{j2\pi m'\Delta f\xi_2}e^{-j2\pi m\Delta f(\xi_2+\tau_p)} \times e^{j2\pi\left(\nu_p+\sum_{q=1}^{Q}a_q(\xi_2+\tau_p+(n-1)T)^q\right)(\xi_2+(n-1)T)} \quad (7).$$

$$h_{k,l}^{(1)}[k',l'] = \sum_{p=0}^{P-1}\frac{\beta_p}{T}(T-\tau_p)e^{j2\pi\nu_p\xi_1}\frac{e^{j2\pi M(\Delta f\xi_1-\frac{l'}{M})}-1}{e^{j2\pi(\Delta f\xi_1-\frac{l'}{M})}-1}\frac{e^{-j2\pi M(\Delta f(\xi_1+\tau_p)-\frac{l}{M})}-1}{e^{-j2\pi(\Delta f(\xi_1+\tau_p)-\frac{l}{M})}-1} \times \sum_{n=0}^{N-1}e^{j2\pi\left(\left(\sum_{q=1}^{Q}a_q(\xi_1+\tau_p+nT)^q\right)(\xi_1+nT)+\nu_p nT-n\frac{k-k'}{N}\right)} \quad (8),$$

$$h_{k,l}^{(2)}[k',l'] = \sum_{p=0}^{P-1}\frac{\beta_p\tau_p}{T}e^{j2\pi\nu_p\xi_2}\frac{e^{j2\pi M(\Delta f\xi_2-\frac{l'}{M})}-1}{e^{j2\pi(\Delta f\xi_2-\frac{l'}{M})}-1} \times \frac{e^{-j2\pi M(\Delta f(\xi_2+\tau_p)-\frac{l}{M})}-1}{e^{-j2\pi(\Delta f(\xi_2+\tau_p)-\frac{l}{M})}-1} \times \sum_{n=1}^{N-1}e^{j2\pi\left(\left(\sum_{q=1}^{Q}a_q(\xi_2+\tau_p+(n-1)T)^q\right)(\xi_2+(n-1)T)+\nu_p(n-1)T-\frac{nk}{N}+\frac{(n-1)k'}{N}\right)} \quad (9).$$

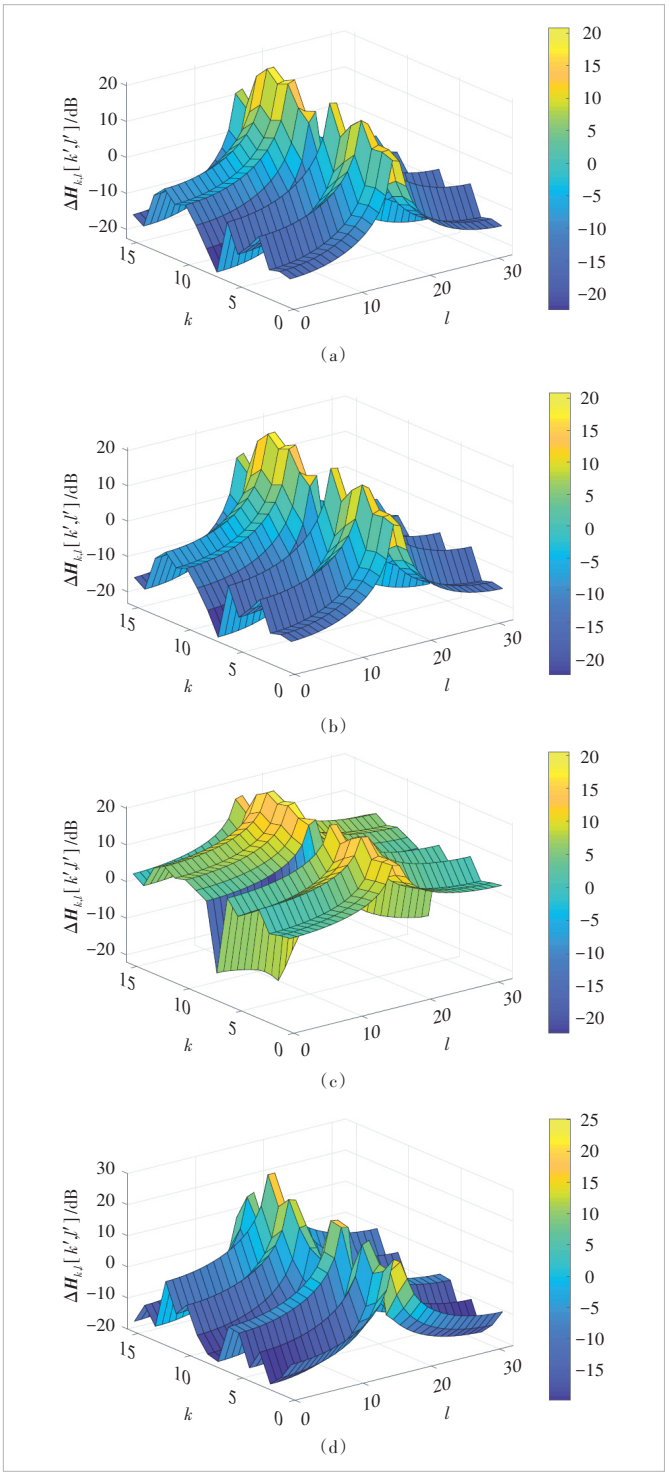
For an explicit illustration of the Doppler rate effect, we mesh the difference of the delay-Doppler domain channel $\Delta H_{k,l}[k',l'] = |h_{k,l}[k',l'] - \hat{h}_{k,l}[k',l']|^2$, given the fixed $k' = \frac{N}{2}$ and $l' = \frac{M}{2}$, where $\hat{h}_{k,l}[k',l']$ is the delay-Doppler domain channel without the Doppler rate influence, i.e., $a_q = 0, \forall q$. The values of N and M are set as 16 and 32, respectively. As the different values of ξ_1 and ξ_2 would have little impact on the characterization of the channel, we set $\xi_1 = \frac{T-\tau_{\max}}{2}$, $\xi_2 = \frac{2T-\tau_{\max}}{2}$.

The highest orders of the Doppler rate in Figs. 1 and 2 are 1



▲ **Figure 1. Difference of the delay–Doppler domain channel matrix $\Delta H_{k,l}[k',l']$ meshes with the highest order of the Doppler rate Q=1, where the values of the a_1 in (a), (b), (c), and (d) are set as 49 Hz/s, 490 Hz/s, 4 900 Hz/s, and 49×10^8 Hz/s, respectively**

and 2, respectively. We can see that the difference of the delay-Doppler domain channel becomes larger as the values of the Doppler rate increase. Compared with Fig. 1, we can see that the variance of the delay-Doppler domain channel is more sophisticated under a higher order of the Doppler rate in Fig. 2. Therefore, it is necessary to design a scheme to estimate and then compensate the effect of the Doppler rate to guarantee reliable communication in high-speed scenarios.



▲ **Figure 2. Difference of the delay-Doppler domain channel matrix $\Delta \boldsymbol{H}_{k,l}[k',l']$ meshes with the highest order of the Doppler rate Q=2, where the values of a_2 are all set as 49×10^8 Hz/s^2; the values of a_1 in (a), (b), (c), and (d) are set as 49 Hz/s, 490 Hz/s, 4 900 Hz/s, and 49×10^8 Hz/s, respectively**

3 Receiver Design

The joint frame structure is designed, where the time-domain linear frequency modulated signal is sent in the first frame to estimate the Doppler rate and then the Doppler rate compensation is performed in the subsequent frames by using the Doppler rate estimate in the first frame.

The diagram of the proposed scheme in the first frame is demonstrated in Fig. 3. We consider a downlink high-mobility transmission scenario where a large-scale antenna array is arranged at the receiver. In addition, the value of the arbitrary order of the Doppler rate is assumed as a constant in the system model. The multi-path channel from the base station to the b-th antenna is expressed as:

$$h_b(t,\tau) = \sum_{p=0}^{P-1} \beta_p e^{j\left(2\pi\left(\nu_p + \sum_{q=1}^{Q} a_q t^q\right)t + \phi_b \cos\theta_p\right)} \delta(\tau - \tau_p) \tag{10},$$

where $b \in \mathcal{I}_B$, and B is the number of the receive antennas. The phase of the b-th antenna is expressed as:

$$\phi_b = \frac{1}{\lambda} 2\pi b\eta, \quad b \in \mathcal{I}_B \tag{11},$$

where λ is the carrier wavelength; $\eta < 0.5\lambda$ is the antenna distance of the uniform linear array (ULA) and is designed to produce only one beam in each receiving beamformer.

The received signal of the b-th antenna is expressed as:

$$r_b(t) = \sum_{p=0}^{P-1} \beta_p e^{j\left(2\pi\left(f_d \cos\theta_p + \sum_{q=1}^{Q} a_q t^q\right)t + \phi_b \cos\theta_p\right)} s(t - \tau_p) + \tilde{z}_b(t) \tag{12},$$

where $\tilde{z}_b(t)$ is the time domain circularly symmetric complex Gaussian (CSCG) noise of the b-th antenna and it follows $\mathcal{CN}(0,\sigma^2)$ at a time instant.

To separate the multi-path effect, the receive beamforming is implemented by a spatial matched filter. The corresponding steering vector of the b-th antenna is designed as follows.

$$\omega_b(\theta) = e^{j\phi_b \cos\theta}, \quad b \in \mathcal{I}_B \tag{13}.$$

After scanning all possible angles, only U branches receive the desired signal. We assume the interested angles are in the set $\boldsymbol{\Phi} = \{\varphi_u | u \in \mathcal{I}_U\}$. In addition, the one-to-one mapping function is defined as $u = \varpi(p)$, $u \in \mathcal{I}_U$, where the index p maps to the identified path u. Therefore, the received signal from the angle φ_u is represented by:

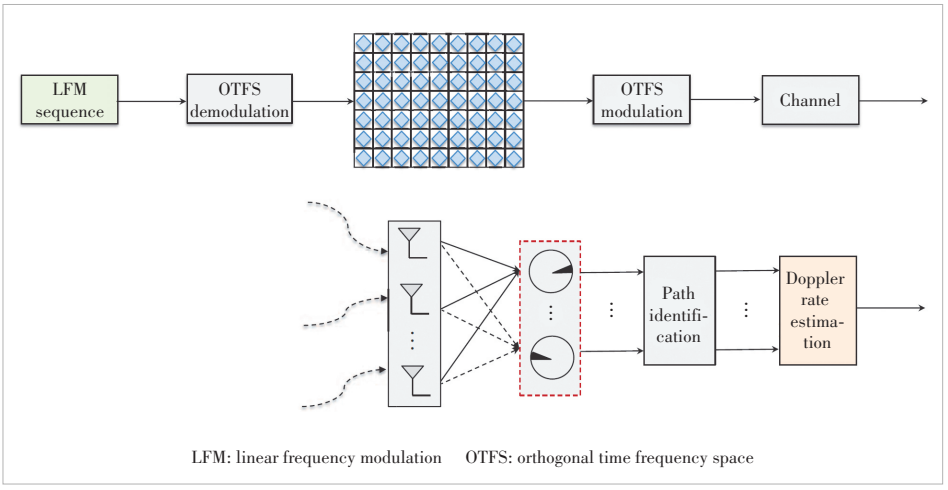


Figure 3. Diagram of the proposed scheme to estimate the Doppler rate in the first frame

$$r_u(t)=\frac{1}{B}\sum_{b=0}^{B-1}\omega_b^*(\varphi_u)r_b(t)=\beta_p e^{j2\pi\left(f_d\cos\theta_p+\sum_{q=1}^{Q}a_q t^q\right)t}s(t-\tau_p)+z_u(t)+\frac{1}{B}\sum_{b=0}^{B-1}\sum_{\varpi(p')\neq u}\beta_{p'}e^{j2\pi\left(f_d\cos\theta_{p'}+\sum_{q=1}^{Q}a_q t^q\right)t}\times e^{j\phi_b\left(\cos\theta_{p'}-\cos\varphi_u\right)} \quad (14),$$

where

$$z_u(t)=\frac{1}{B}\sum_{b=0}^{B-1}\omega_b^*(\varphi_u)\tilde{z}_b(t),\quad u\in\mathcal{I}_U \quad (15).$$

With the arrangement of a large-scale antenna array, the interference of the identified branch can be ignored and was also proved in our previous work[11]. Then the received signal of the u-th identified branch can be expressed as:

$$r_u(t)\approx\beta_u e^{j2\pi\left(\nu_u+\sum_{q=1}^{Q}a_q t^q\right)t}s(t-\tau_u)+z_u(t) \quad (16),$$

where $\beta_u=\beta_p$, $\tau_u=\tau_p$, $\nu_u=f_{\mathrm{d}}\cos\varphi_u$, and $u=\varpi(p)$ is the Doppler shift of the u-th identified branch.

4 Doppler Rate Estimation

In this section, the proposed Doppler rate estimation scheme is introduced to the system with the first-order and the second-order Doppler rate. Then the generalized Doppler rate estimation scheme that applies to the system with an arbitrary-order Doppler rate is illustrated. Finally, the Doppler rate mitigation scheme through precoding is demonstrated.

4.1 First-Order Doppler Rate

For the system only with the first-order Doppler rate, i.e., $Q=1$, the received samples of the u-th identified branch can be represented by:

$$r_u(n)=\beta_u e^{j2\pi\left(\frac{k_u}{MN}n+\frac{a_1}{(M\Delta f)^2}n^2\right)}s(n-l_u) \quad (17),$$

where $k_u=\nu_u NT\in\mathbb{R}$ and $l_u=\lfloor\tau_u M\Delta f+0.5\rfloor$. After the ISFFT and the Heisenberg transform, the time domain linear frequency modulated sequence $s(n)$ is sent with the length $N_0=MN-1$. The estimation of the Doppler rate conducts as follows.

1) Calculate the instantaneous autocorrelation of the $r_u^*(n)$:

$$A_r(n)=r_u^*(n)r_u(n+\eta_0)=|\beta_u|^2 e^{j2\pi\left(\frac{k_u}{MN}\eta_0+\frac{a_1}{(M\Delta f)^2}(d_0^2+2\eta_0 n)\right)} \quad (18),$$

where $\eta_0\in(0,N_0)$ is a constant, $n\in[-M_0,M_0-\eta_0]$, and $M_0=(N_0-1)/2$.

2) Calculate the fourth-order moment $F_r(\eta_1)$:

$$F_r(\eta_1)=\sum_{n=N_1}^{N_2}A_r^*(n)A_r(n+\eta_1)=|\beta_u|^4 e^{j2\pi\frac{2a_1\eta_0\eta_1}{(M\Delta f)^2}} \quad (19),$$

where $\eta_1\in[-(N_0-\eta_0-1),N_0-\eta_0-1]$, $N_1=\max\{-M_0,-M_0-\eta_1\}$, and $N_2=\min\{M_0-\eta_0,M_0-\eta_0-\eta_1\}$.

3) Calculate the forth-order moment of $s(n)$ to obtain $F_s(\eta_1)$. The calculation of the forth-order moment of $s(n)$ follows Eq. (19).

4) Calculate the ratio between $F_r(\eta_1)$ and $F_s(\eta_1)$:

$$\xi(\eta_1)=\begin{cases}\dfrac{F_r(\eta_1)}{F_s(\eta_1)}, & F_s(\eta_1)\neq 0 \text{ and } \eta_1\neq 0\\ \dfrac{\xi(\eta_1-1)+\xi(\eta_1+1)}{2}, & \text{otherwise}\end{cases} \quad (20).$$

5) Perform fast Fourier transform on $\xi(\eta_1)$:

$$\Xi(\bar{l})=\sum_{\eta_1=-(N_0-\eta_0-1)}^{N_0-\eta_0-1}\xi(\eta_1)e^{-j\frac{2\pi\eta_1\bar{l}}{2(N_0-\eta_0)-1}},\quad \bar{l}\in[-(N_0-\eta_0-1),(N_0-\eta_0-1)] \quad (21).$$

6) Maximize $\Xi(\bar{l})$. Find the value of $\bar{l}$ that maximizes $\Xi(\bar{l})$ and then estimate the first-order Doppler rate a_1 as:

$$\hat{a}_1=\frac{(M\Delta f)^2}{2\eta_0(2N_0-2\eta_0-1)}\operatorname*{argmax}_{\bar{l}}|\Xi(\bar{l})| \quad (22).$$

4.2 Second-Order Doppler Rate

For the system with the second-order Doppler rate, i.e. $Q=2$,

the received samples can be represented by

$$r_u(n) = \beta_u e^{j2\pi\left(\frac{k_u}{MN}n + \frac{a_1}{(M\Delta f)^2}n^2 + \frac{a_2}{(M\Delta f)^3}n^3\right)} s(n - l_u) \tag{23}$$

To cancel the influence of Doppler rates, the designed scheme first estimates the second-order Doppler rate. Then, the effect of the second-order Doppler rate is removed from the received samples. Next, the first-order Doppler rate is estimated and then removed. Based on the estimation of the first-order Doppler rate, the second-order Doppler rate is calculated as follows.

1) Calculate the eighth-order moment $E_r(\eta_2)$:

$$E_r(\eta_2) = F_r(\eta_1)F_r^*(\eta_1 + \eta_2) = \left|\beta_u\right|^8 e^{j2\pi\frac{6a_2\eta_0\eta_1\eta_2}{(M\Delta f)^3}} \tag{24}$$

where $\eta_2 \in [-(N_0 - \eta_0 - \eta_1 - 1), N_0 - \eta_0 - \eta_1 - 1]$.

2) Calculate the eighth-order moment $E_s(\eta_2)$. The calculation of the eighth-order moment of $s(n)$ follows Eq. (27).

3) Calculate the ratio between $E_r(\eta_2)$ and $E_s(\eta_2)$:

$$\xi(\eta_2) = \begin{cases} \dfrac{E_r(\eta_2)}{E_s(\eta_2)}, & E_s(\eta_1) \neq 0 \text{ and } \eta_2 \neq 0 \\ \dfrac{\xi(\eta_2 - 1) + \xi(\eta_2 + 1)}{2}, & \text{otherwise} \end{cases} \tag{25}$$

4) Perform the fast Fourier transform on $\xi(\eta_2)$:

$$\Xi(\tilde{l}) = \sum_{\eta_2 = -(N_0 - \eta_0 - \eta_1 - 1)}^{N_0 - \eta_0 - \eta_1 - 1} \xi(\eta_2) e^{-j\frac{2\pi\eta_2\tilde{l}}{2(N_0 - \eta_0 - \eta_1) - 1}},$$
$$\tilde{l} \in [-(N_0 - \eta_0 - \eta_1 - 1), (N_0 - \eta_0 - \eta_1 - 1)] \tag{26}$$

5) Maximize $|\Xi(\tilde{l})|$. Find the value of $\tilde{l}$ that maximizes $|\Xi(\tilde{l})|$ and then estimate the second-order Doppler rate a_2 as:

$$\hat{a}_2 = \frac{(M\Delta f)^3}{6\eta_0\eta_1(2N_0 - 2\eta_0 - 2\eta_1 - 1)} \underset{\tilde{l}}{\operatorname{argmax}} |\Xi(\tilde{l})| \tag{27}$$

4.3 Extension to Higher-Order Doppler Rate

We can extend the proposed Doppler rate estimation scheme to a system with an arbitrary-order Doppler rate. For a system with Q-th order Doppler rate, the received samples can be expressed as:

$$r_u(n) = \beta_u e^{j2\pi\left(\frac{k_u}{MN}n + \sum_{q=1}^{Q} a_q\left(\frac{n}{M\Delta f}\right)^{q+1}\right)} s(n - l_u) \tag{28}$$

The estimation of the Doppler rate is conducted with a successive interference pattern as follows. The 2^{Q+1}-order moment of the received samples is calculated first. Then the ratio between the 2^{Q+1}-order moment of the received samples and the 2^{Q+1}-order moment of the sent samples is calculated. Next, the fast Fourier transform is utilized to transform the ratio into the frequency domain and obtain the spectrum of the moment. Finally, the spectrum is maximized and the corresponding estimation of the Q-th order Doppler rate a_Q is calculated. Once the Q-th order Doppler rate is estimated, it is removed from the received samples, and the $(Q-1)$-th order Doppler rate is calculated and then cancelled. The estimation and the compensation processes continue until the first-order Doppler rate is mitigated. The proposed Doppler rate scheme is unbiased, which can be proved based on the fact that the noise is zero mean CSCG and the Fourier transform does not change the mean value.

4.4 Precoding Scheme

For the frames that transmit the information-bearing symbols, precoding is performed in the delay-Doppler domain to mitigate the effect of the Doppler rate.

From Ref. [4], the time-domain transmitted symbol vector $\boldsymbol{s}$ can be expressed as:

$$\boldsymbol{s} = (\boldsymbol{F}_N^{\mathrm{H}} \otimes \boldsymbol{I}_M)\boldsymbol{x}_D \tag{29}$$

where $\boldsymbol{x}_D$ is the delay-Doppler domain transmit vector and the operation $\otimes$ denotes the Kronecker product. To mitigate the effect of the Doppler rate, time-domain precoding is carried out as:

$$\boldsymbol{P}\boldsymbol{s} = \boldsymbol{P}(\boldsymbol{F}_N^{\mathrm{H}} \otimes \boldsymbol{I}_M)\boldsymbol{x}_D \tag{30}$$

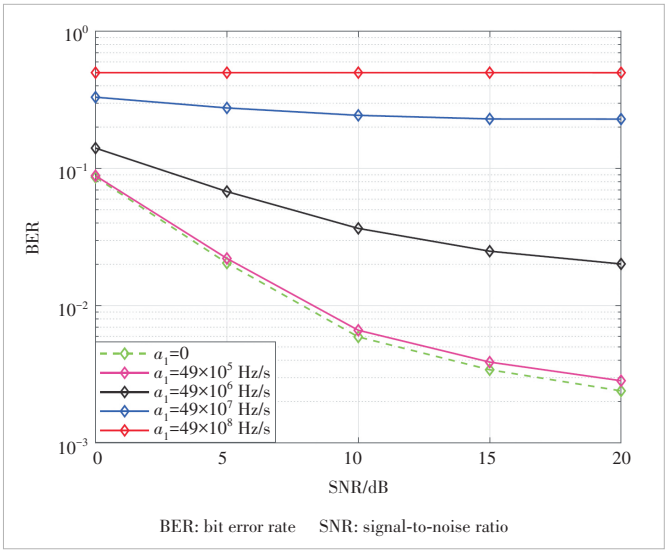
where $\boldsymbol{P} = \mathrm{diag}\{\boldsymbol{p}\}$ and the n-th element of the vector is $p_n = e^{-j2\pi\left(\sum_{q=1}^{Q}\hat{a}_q\left(\frac{n}{M\Delta f}\right)^{q+1}\right)}$, $n \in \mathcal{I}_{MN}$. Since the information-bearing symbols are transmitted in the delay-Doppler domain, the precoding matrix is designed in the delay-Doppler domain as $\boldsymbol{P}(\boldsymbol{F}_N^{\mathrm{H}} \otimes \boldsymbol{I}_M)$.

5 Simulation Results

In this section, we evaluate the performance from the normalized mean squared error (NMSE) of the Doppler rate estimation and the bit error rate (BER) of the proposed scheme. The NMSE of the q-th order Doppler rate is defined as $10\log_{10}\left(|a_q - \hat{a}_q|^2/a_q^2\right)$. We set $N = 32$, $M = 64$, $\Delta f = 150$ kHz, the moving speed $v = 1$ Ma, the carrier frequency $f_c = 4$ GHz, and the modulation scheme is 4 Quadrature Amplitude Modulation (QAM). The delay indices of the channel is $[0,1,2,3,4,5]$ and the power of each tap is uniformly distributed. The Doppler shift of the channel is generated by $f_d \cos\theta_p$ where the angle of arrival (AoA) of each path θ_p is independently uniformly distributed in $[0,2\pi)$. Furthermore, the channel response of all the antenna elements are normalized as 1. Moreover, the receive beamforming is designed with an interval of one degree. For the error performance, the channel estimation in Ref. [11] and the data detection scheme in Ref. [12] are adopted. The value of the Doppler rate is set as large as possible and such setting is suitable especially under the take-off and the landing process of high-speed aircrafts.

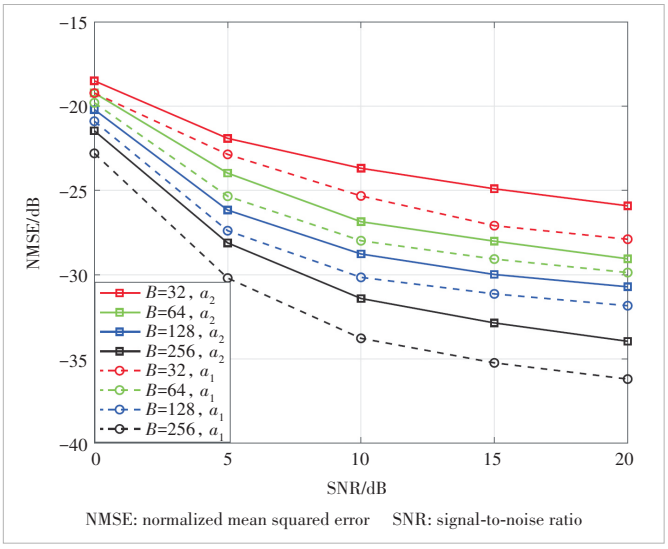
In Fig. 4, we demonstrate the error performance by introducing the different values of the Doppler rate. In addition, the BER is evaluated without the estimation and compensation of the Doppler rate. We can obtain that the error performance when $a_1 \leqslant 49 \times 10^5$ Hz/s is nearly the same as that without the introduction of the Doppler rate. However, the error performance deteriorates sharply when $a_1 \geqslant 49 \times 10^5$ Hz/s. Especially, the system cannot even normally work when $a_1 = 49 \times 10^8$ Hz/s. The reason can be explained by comparing the maximum Doppler shift of one OTFS frame Δf and the Doppler shift increment from the Doppler rate $\Delta D = a_1 NT$. When $a_1 = 49 \times 10^5$ Hz/s, the increment of the Doppler shift is $\Delta D = \frac{49 \times 10^5 \times 32}{150 \times 10^3} \approx 1\,045$ Hz $< 15 \times 10^3$ Hz= Δf. Though without the estimation and the compensation of the Doppler rate, the effect of the Doppler rate can be mitigated in the channel estimation. Thus, the error performance can keep the same as that without the Doppler rate. However, the increment of the Doppler shift can arrive at the value $\Delta D = \frac{49 \times 10^6 \times 32}{150 \times 10^3} \approx 1\,045$ kHz> Δf. The Doppler rate cannot be mitigated from the channel estimation deteriorating the error performance. Therefore, the estimation and the compensation of the Doppler rate are necessary to guarantee the reliable communication under such setting.

In Fig. 5, we evaluate the Doppler rate estimation accuracy under $Q = 2$, $a_2 = 49 \times 10^{12}$ Hz/s^2, and $a_1 = 49 \times 10^8$ Hz/s. We can see that the estimation accuracy of both the second order Doppler rate and the first order Doppler rate improves with the increasing number of the receive antennas. In addition, the estimation accuracy of the first order Doppler rate under $Q = 1$ outperforms that under $Q = 2$. This phenomenon is caused by the proposed successive interference cancellation pattern for the high order Doppler rate.

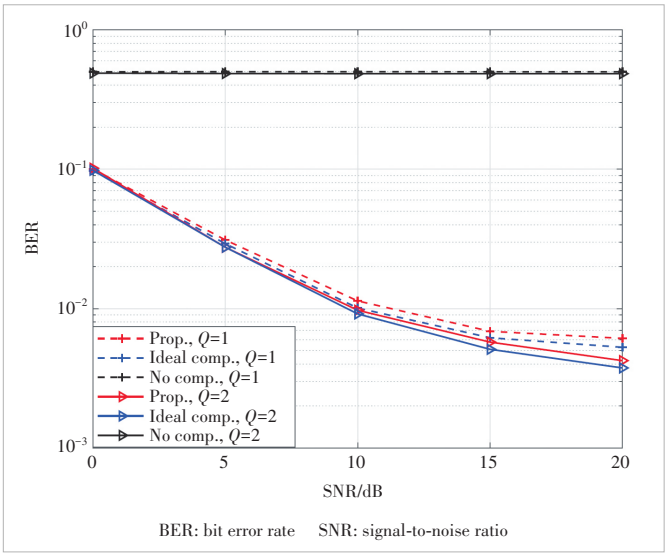


▲ **Figure 4. BER is evaluated under the different values of the Doppler rate and without the Doppler rate effect compensation, where the results demonstrate that the error performance deteriorates as the Doppler rate increases**

In Fig. 6, we demonstrate the error performance of the proposed transceiver under $Q = 1$ and $Q = 2$. In addition, $\eta_0 = 400$ and $a_1 = 49 \times 10^8$ Hz/s for $Q = 1$; $\eta_0 = 800$, $\eta_1 = 400$, $a_2 = 49 \times 10^{12}$ Hz/s^2, and $a_1 = 49 \times 10^8$ Hz/s for $Q = 2$. We can see that the proposed transceiver can achieve the same error performance as that with the perfect Doppler rate compensation, indicating the proposed transceiver can effectively mitigate the Doppler rate effect. In addition, the system cannot even normally work without compensating the Doppler rate, which il-



▲ **Figure 5. NMSE of the Doppler rate is evaluated under the four values of the receive antenna, namely 32, 64, 128, and 256. The highest order of the Doppler rate is 2, i.e., $Q = 2$, $a_1 = 49 \times 10^8$ Hz/s and $a_2 = 49 \times 10^{12}$ Hz/s^2. We can see that the performance of the proposed transceiver improves with the increasing number of the receive antenna**



▲ **Figure 6. BER is evaluated under the three schemes and the two values of the highest order of the Doppler rate. We can see that the proposed scheme can achieve nearly the same performance as the perfect Doppler rate compensation under both the first order and the second order Doppler rate conditions**

lustrates the significance of estimating the Doppler rate effect.

6 Conclusions

In this paper, we first introduce the effect of the Doppler rate in the OTFS system and derive the delay-Doppler domain input-output relation. Then the Doppler rate effect is characterized by utilizing the first mean value theorem for definite integrals to avoid the complicated integrals. Aiming at mitigating the Doppler rate effect, the joint frame transceiver scheme, where the Doppler rate is estimated in the first frame and then the effect is removed in the subsequent data frames, is designed by arranging a large-scale antenna array at the receiver. Simulation results demonstrate the efficiency of the proposed scheme.

References

[1] HADANI R, RAKIB S, TSATSANIS M, et al. Orthogonal time frequency space modulation [C]//IEEE Wireless Communications and Networking Conference (WCNC). IEEE, 2017: 1 – 6. DOI: 10.1109/WCNC.2017.7925924
[2] HADANI R, MONK A. OTFS: a new generation of modulation addressing the challenges of 5G [EB/OL]. (2018-02-17) [2023-06-15]. http://arxiv. org/abs/1802.02623
[3] SURABHI G D, AUGUSTINE R M, CHOCKALINGAM A. On the diversity of uncoded OTFS modulation in doubly-dispersive channels [J]. IEEE transactions on wireless communications, 2019, 18(6): 3049 – 3063. DOI: 10.1109/TWC.2019.2909205
[4] RAVITEJA P, HONG Y, VITERBO E, et al. Practical pulse-shaping waveforms for reduced-cyclic-prefix OTFS [J]. IEEE transactions on vehicular technology, 2019, 68(1): 957 – 961. DOI: 10.1109/TVT.2018.2878891
[5] HUANG P H, LIAO G S, YANG Z W, et al. Long-time coherent integration for weak maneuvering target detection and high-order motion parameter estimation based on keystone transform [J]. IEEE transactions on signal processing, 2016, 64(15): 4013 – 4026. DOI: 10.1109/TSP.2016.2558161
[6] HUANG P H, LIAO G S, YANG Z W, et al. Ground maneuvering target imaging and high-order motion parameter estimation based on second-order keystone and generalized hough-HAF transform [J]. IEEE transactions on geoscience and remote sensing, 2017, 55(1): 320 – 335. DOI: 10.1109/TGRS.2016.2606436
[7] ZHU C Y, LI X P, SHI L, et al. A new fast Doppler shift and Doppler rate joint acquisition method for hypersonic vehicle communications [C]//Proceedings of International Symposium on Antennas and Propagation (ISAP). IEEE, 2018: 1 – 2
[8] WAQAS A, LECHNER G, NGUYEN K, et al. Particle filter for joint carrier phase, Doppler shift and Doppler rate estimation and data detection [C]//Proceedings of IEEE Latin-American Conference on Communications (LATINCOM). IEEE, 2021: 1 – 6. DOI: 10.1109/LATINCOM53176.2021.9647810
[9] LIU Y S, ZHANG S, GAO F F, et al. Uplink-aided high mobility downlink channel estimation over massive MIMO-OTFS system [J]. IEEE journal on selected areas in communications, 2020, 38(9): 1994 – 2009. DOI: 10.1109/JSAC.2020.3000884
[10] DONG K Y, SHI J, WANG Z Y, et al. Performance analysis of orthogonal time frequency space modulation under time-varying Doppler channels [C]// The 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC). IEEE, 2023: 1 – 6. DOI: 10.1109/PIMRC56721.2023.10293809
[11] SHAN Y R, WANG F G. Low-complexity and low-overhead receiver for OTFS via large-scale antenna array [J]. IEEE transactions on vehicular technology, 2021, 70(6): 5703 – 5718. DOI: 10.1109/TVT.2021.3072667
[12] RAVITEJA P, PHAN K T, HONG Y, et al. Interference cancellation and iterative detection for orthogonal time frequency space modulation [J]. IEEE transactions on wireless communications, 2018, 17(10): 6501 – 6515. DOI: 10.1109/TWC.2018.2860011

Biographies

SHAN Yaru received her BE degree from the School of Electronic and Information Engineering, Beijing Information Science and Technology University, China in 2018. She is currently pursuing her PhD degree with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, China. Her current research interests include signal processing, orthogonal time frequency space, signal detection in high-speed scenarios, and integrated sensing and communication.

WANG Fanggang (fgwang@bjtu.edu.cn) received his BE and PhD degrees from the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, China in 2005 and 2010, respectively. He was a post-doctoral fellow with the Institute of Network Coding, The Chinese University of Hong Kong, China from 2010 to 2012. He was a visiting scholar with the Singapore University of Technology and Design in 2014, and with the Massachusetts Institute of Technology, USA from 2015 to 2016. He is currently a professor with the State Key Laboratory of Advanced Rail Autonomous Operation, Frontiers Science Center for Smart High-Speed Railway System, School of Electronics and Information Engineering, Beijing Jiaotong University, China. His research interests include wireless communications, signal processing, and information theory. He served as a technical program committee member for several conferences. He served as an editor for *IEEE Communications Letters*.

HAO Yaxing received his BE degree from the School of Electronic and Information Engineering, Beijing Jiaotong University, China in 2021, where he is currently pursuing his PhD degree with the State Key Laboratory of Rail Traffic Control and Safety. His current research interests include signal processing, orthogonal time frequency space, and signal detection in high-speed scenario.

HUA Jian received his MS degree from Harbin Engineering University, China, and now works at ZTE Corporation as an intermediate engineer. His research interests include phase noise model and compensation scheme design, waveform modulation and other technologies in terahertz communication scenarios.

XIN Yu graduated with a PhD degree from Beijing University of Posts and Telecommunications, China in 2003. He currently works at ZTE Corporation as a senior engineer and senior expert in technology pre-research, specializing in wireless communication technology. He first put forward the FB-OFDM and GFBOFDM waveform schemes and has published dozens of papers on waveform research. He is currently responsible for pre-research on candidate new waveforms for 6G.

The 1st Youth Expert Committee
for Promoting Industry-University-Institute Cooperation

Director	**CHEN Wei,** Beijing Jiaotong University
Deputy Director	**QIN Xiaoqi,** Beijing University of Posts and Telecommunications
	LU Dan, ZTE Corporation

Members (Surname in Alphabetical Order)

CAO Jin	Xidian University
CHEN Li	University of Science and Technology of China
CHEN Qimei	Wuhan University
CHEN Shuyi	Harbin Institute of Technology
CHEN Siheng	Shanghai Jiao Tong University
CHEN Wei	Beijing Jiaotong University
GUAN Ke	Beijing Jiaotong University
HAN Kaifeng	China Academy of Information and Communications Technology
HE Zi	Nanjing University of Science and Technology
HOU Tianwei	Beijing Jiaotong University
HU Jie	University of Electronic Science and Technology of China
HUANG Chen	Purple Mountain Laboratories
LI Ang	Xi'an Jiaotong University
LIU Chunsen	Fudan University
LIU Fan	Southeast University
LIU Junyu	Xidian University
LU Dan	ZTE Corporation
LU Youyou	Tsinghua University
NING Zhaolong	Chongqing University of Posts and Telecommunications
QI Liang	Shanghai Jiao Tong University
QIN Xiaoqi	Beijing University of Posts and Telecommunications
QIN Zhijin	Tsinghua University
SHI Yinghuan	Nanjing University
TANG Wankai	Southeast Univeristy
WANG Jingjing	Beihang University
WANG Xinggang	Huazhong University of Science and Technology
WANG Yongqiang	Tianjin University
WEN Miaowen	South China University of Technology
WU Qingqing	Shanghai Jiao Tong University
WU Yongpeng	Shanghai Jiao Tong University
XIA Wenchao	Nanjing University of Posts and Telecommunications
XU Mengwei	Beijing University of Posts and Telecommunications
XU Tianheng	Shanghai Advanced Research Institute, Chinese Academy of Sciences
YANG Chuanchuan	Peking University
YIN Haifan	Huazhong University of Science and Technology
YU Jihong	Beijing Institute of Technology
ZHANG Jiao	Beijing University of Posts and Telecommunications
ZHANG Yuchao	Beijing University of Posts and Telecommunications
ZHANG Jiayi	Beijing Jiaotong University
ZHAO Yuda	Zhejiang University
ZHAO Zhongyuan	Beijing University of Posts and Telecommunications
ZHOU Yi	Southwest Jiaotong University
ZHU Bingcheng	Southeast University

ZTE COMMUNICATIONS
中兴通讯技术(英文版)

ZTE Communications has been indexed in the following databases:

- Abstract Journal
- China Science and Technology Journal Database
- Chinese Journal Fulltext Databases
- Index Copernicus
- Scopus
- Ulrich's Periodicals Directory
- Wanfang Data
- WJCI 2021–2024

ZTE COMMUNICATIONS
Vol. 23 No. 1 (Issue 90)
Quarterly
First Issue Published in 2003

Supervised by:
Anhui Publishing Group

Sponsored by:
Time Publishing and Media Co., Ltd.
Shenzhen Guangyu Aerospace Industry Co., Ltd.

Published by:
Anhui Science & Technology Publishing House

Edited and Circulated (Home and Abroad) by:
Magazine House of ZTE Communications

Editorial Correspondence:
Add: 12F Kaixuan Building, 329 Jinzhai Road,
Hefei 230061, P. R. China
Tel: +86-551-65533356
Email: magazine@zte.com.cn
Website: http://zte.magtechjournal.com

Annual Subscription: RMB 120
Printed by:
Hefei Tiancai Color Printing Company
Publication Date: March 25, 2025

China Standard Serial Number: ISSN 1673–5188 / CN 34–1294/TN