基于深度生成模型的视觉模式表示与编码

Visual Pattern Representation and Coding Based on Deep Generative Models

郭怡琳/GUO Yilin[1]，常建慧/CHANG Jianhui[2]，
黄成/HUANG Cheng[3]，马思伟/MA Siwei[2,4]
（1. 北京大学深圳研究生院，中国 深圳 518055；
2. 北京大学，中国 北京 100871；
3. 中兴通讯股份有限公司，中国 深圳 518057；
4. 鹏城实验室，中国 深圳 518057）
(1. Peking University Shenzhen Graduate School, Shenzhen 518055, China;
2. Peking University, Beijing 100871, China;
3. ZTE Corporation, Shenzhen 518057, China;
4. Pengcheng Laboratory, Shenzhen 518057, China)

DOI：10.12142/ZTETJ.2024S1009
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20240726.1421.002.html
网络出版日期：2024-07-29
收稿日期：2023-11-25

摘要：认为早期智能编码方法的性能受限于手工设计的方案，当前基于神经网络的编码方法可解释性不足，不利于后续面向人机视觉的分析与交互。受生成模型的启发，生成式编码方法通过构建生成模型来实现图像和视频的压缩和合成，获得可解释的紧凑视觉表示并生成符合图像先验分布的高视觉质量内容。其中概念图像编码与概念视频编码利用生成模型强大的样本生成能力与紧凑层次视觉表示模型，实现了编码性能更优的图像与视频编码；跨模态语义编码对图像与文本域进行跨模态转换与编码，保持可解释的同时实现上千倍的超高压缩比与令人满意的重构结果。

关键词：智能视频编码；生成式编码；跨模态压缩；概念编码

Abstract: The performance of early intelligent encoding methods was limited by manually designed solutions, while current neural network-based encoding methods lack interpretability, which hinders subsequent analysis and interaction between humans and machine vision. Inspired by generative models, the generative encoding methods aim to achieve compression and synthesis of images and videos by constructing efficient generative models, obtaining interpretable compact visual representations, and synthesizing high-quality visual content that conforms to the prior distribution of images. Among them, conceptual image encoding and conceptual video encoding leverage the powerful sample generation capability and compact hierarchical visual representation models of generative models, resulting in superior encoding performance for images and videos. Cross-modal semantic coding, on the other hand, enables cross-modal transformation and coding between the image and text domains while maintaining interpretability, achieving ultra-high compression ratios of thousands of times and satisfactory reconstruction results.

Keywords: intelligent video encoding; generative encoding; cross-modal compression; conceptual coding

引用格式：郭怡琳, 常建慧, 黄成, 等. 基于深度生成模型的视觉模式表示与编码 [J]. 中兴通讯技术, 2024, 30(S1): 60-66. DOI: 10.12142/ZTETJ.2024S1009
Citation: GUO Y L, CHANG J H, HUANG C, et al. Visual pattern representation and coding based on deep generative models [J]. ZTE technology journal, 2024, 30(S1): 60-66. DOI: 10.12142/ZTETJ.2024S1009

智能视频编码技术可以追溯到20世纪80年代末[1]，旨在使用知识和语义来设计紧凑的视觉内容表示模型和方法，以便在不同的粒度级别（如块、网格、区域和对象）上对视觉信息进行结构化描述。然而，早期的研究主要针对特定信号源，并基于专家知识进行人工设计编码方案，导致编码性能与应用范围受限。随着移动设备、监控摄像头和其他视频采集设备的大量增加，视频数据量显著增加。在大数据时代，图像和视频处理需要更高效和智能的编码技术。在过去几年中，神经网络在图像和视频理解、处理和压缩等多个

基金项目：国家自然科学基金项目（62025101）；鹏城实验室重大攻关项目（PCL2024A02）

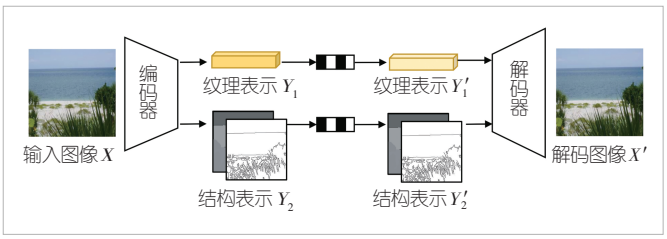
领域展示了巨大的潜力。在压缩任务中，神经网络中的参数可以基于大量图像视频样本进行训练，通过非线性变换将输入的像素数据映射到更紧凑的潜在表示[2]。此外，通过熵估计模型获得的可微码率约束，协同失真约束能够实现自适应率失真优化，能够有效减轻模型对手动设计模块的依赖。得益于这些优良的特性，基于深度学习的编码可以在图像和视频编码领域发挥出更大的潜力。但现有基于深度学习的编码方法将信号数据压缩为不可解释的紧凑潜在表示，这种机制不仅对面向人类视觉的分析与交互不够友好，也无法高效协助下游的机器分析任务。

近期生成模型的快速发展为图像和视频压缩的研究带来了新的启示。生成模型可以对联合概率分布进行建模，从统计的角度表示数据分布情况，并通过控制采样过程生成符合目标分布的样本。生成对抗网络[3]（GAN）、变分自编码器[4]（VAE）和深度扩散模型[5]（DPM）等生成模型在多模态交互和可控生成视觉内容方面展现出强大的能力。本文认为，新兴的生成式编码方法可以通过构建生成模型来实现图像和视频的压缩和合成，通过学习图像和视频数据的先验分布和跨域映射来挖掘样本的潜在特征和数据间的相关性，获得可解释的紧凑视觉表示，从而帮助人类理解和机器分析。此外，在数据信号严重损失的情况下，生成式编码允许生成符合图像先验分布和人眼视觉感知的纹理，从而在带宽受限的条件下保持高质量的视觉重建。作为新一代智能编码的关键技术，当前生成式编码的研究致力于构建视觉内容的概念表示模型，将图像、视频等内容编码成可解释的高层结构、纹理和语义表示，提出了概念图像编码[6-8]、概念视频编码[9]和跨模态语义编码[10]等方法,能够保持可解释性的同时在上千倍的超高压缩比下获得主观感知良好的重构结果。

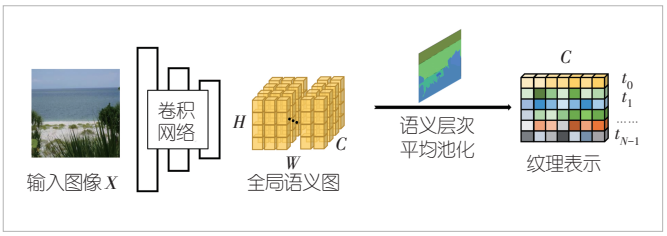
1 概念图像编码

人类视觉系统（HVS）[11]通过处理多方面信息并将其整合为抽象的高级概念（如结构、纹理和语义）来感知视觉内容，从而形成后续认知过程的基础[12]。受HVS启发，概念编码旨在将图像编码成紧凑的、高维的、可解释的表示，以获得高视觉质量的重构与更高效、更便于分析的压缩架构。在解码端，多层解码表征融合并以深度生成的方式合成目标图像。概念编码面临的主要挑战包括如何实现高效的表征分离，以及如何设计有效的生成模型以实现高视觉质量的重建。GREGOR 等[13]引入了带注意力的卷积深度递归输入器（DRAW）[14]，扩展了VAE[4]，使用RNN作为编码器和解码器，将图像转换成一系列越来越细节的表示。然而，其学习到的图像表征的可解释性仍然不足，且仅适用于小分辨率的数据集。神经视频压缩也受到类似的限制，典型的视频压缩方法[15]与图像压缩方法[13]共享相同的VAE架构，并将原始序列转换为低维表示。然而，学习到的视频表征的可解释性仍然缺乏探索。因此，我们在压缩过程中引入结构信息或高级语义信息等可解释表征，以增强图像和视频表征的可解释性。

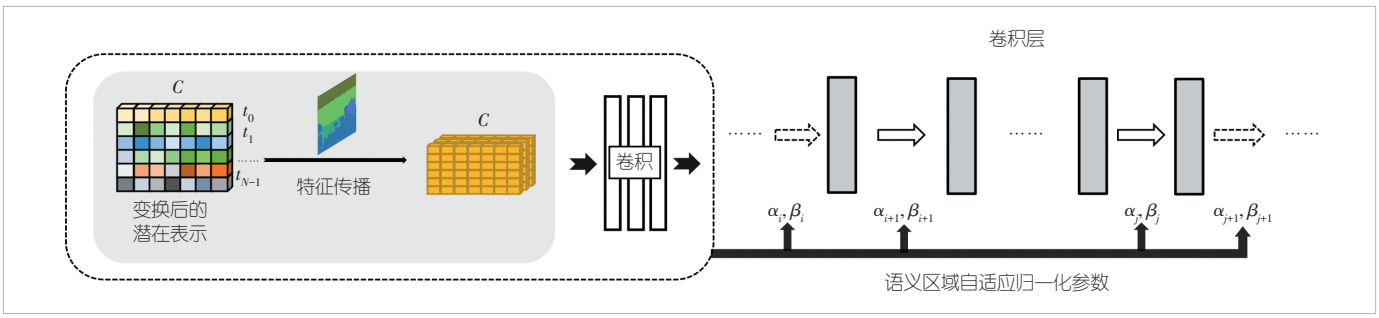
概念编码的突破性进展在于将图像编码为互补的两层视觉概念表示[6-7]。如图1所示，图像被分解为结构和纹理表示两部分，其中纹理表示通过图2展示的纹理建模过程得到，在解码端由图3所示的模块实现图像合成。图4展示了该研究中提出的特征域空间中结构和纹理表征分离。结构层由边



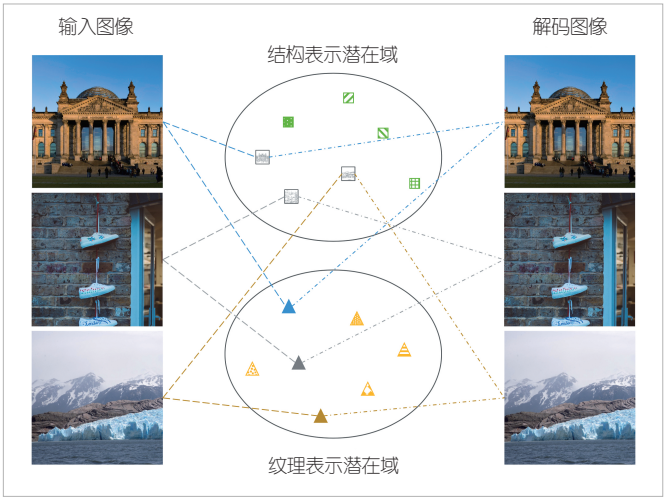
▲图1 典型的概念图像编码框架



▲图2 典型的概念图像编码纹理建模过程



▲图3 典型的概念图像编码合成过程



▲图4 典型的概念图像编码图示

缘图表示，纹理层由变分自编码器提取的低维潜在变量表示。为了从压缩的分层特征中重建原始图像，分层融合生成模型被设计出来[8]，其中纹理层和结构层通过自适应实例归一化（AdaIN）融合引导生成过程。大量实验表明[7]我们提出的概念压缩框架可实现在极低的比特率（<0.1 bpp）下保持较高的视觉重构质量，并通过结构编辑与风格变换以及关键点检测实验证明了概念编码在内容控制和分析任务上的优势。然而，仅使用一组隐变量对整个图像的复杂纹理进行建模是一项巨大的挑战；此外，如何为视觉表征建立有效的熵模型并联合率失真优化，尚未进行过探讨。文献[8]提出了用于语义先验引导的纹理建模，探索了语义粒度纹理表示建模与压缩，以实现高质量的图像合成和可观的编码效率。此外，文献[8]还提出了一种跨通道熵模型，用于联合纹理表示压缩和重建优化。文献[16]进一步引入了结构建模，提出了一种一致性对比学习方法，通过将纹理表示空间与源像素空间对齐来优化纹理表示空间，从而获得更高的压缩性能。与特定应用领域中最先进的多功能视频编码（VVC）相比，文献[8,16]中提出的方法以超低比特率（<0.1 bpp）实现了卓越的视觉重构质量。

由于概念编码方法追求在极低的比特率下取得视觉上令人信服的重构结果，在评价时除了用户打分之外，通常会选择可学习感知图像块相似度（LPIPS）指标[17]作为量化的感知失真度量指标。在之前建立的基准[18]中，该指标已被证明与人类视觉感知而非信号保真度高度相关。为了进行性能比较，表1列出了VVC、典型的端到端学习图像编码方法（E2E）[19]、分层概念编码方法（LCIC）[7]和语义先验模型（SPM）[8]在FFHQ[20]和ADE20K[21]户外测试集中低比特率范围内的率失真性能。结果表明，与基于信号的压缩方法相比，

▼表1 VVC、E2E[19]以及概念编码方法LCIC[7]和SPM[8]在FFHQ和ADE20K户外测试集上的定量结果（LPIPS为失真指标）

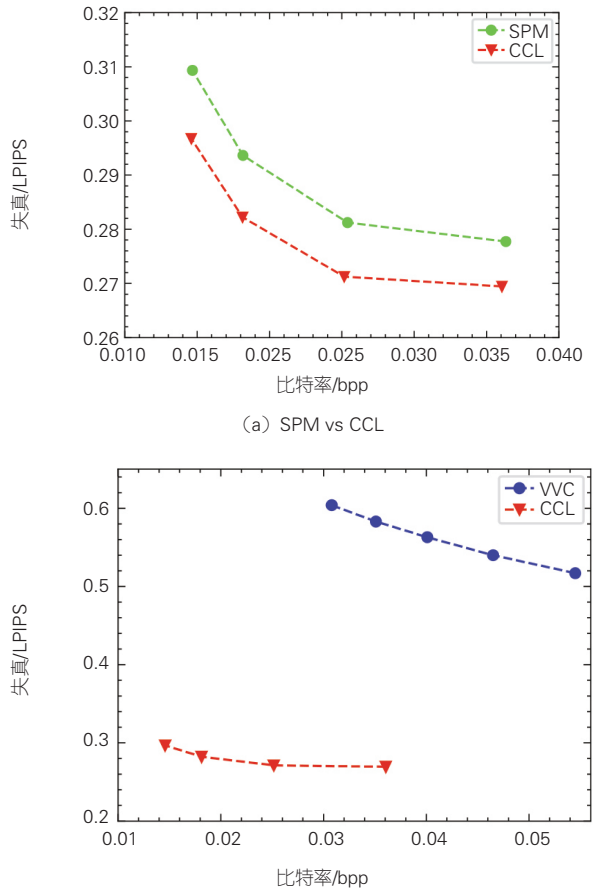
性能	数据集			
	FFHQ		ADE20K	
	比特率/（bpp）	LPIPS（10^{-2}↓）	比特率/（bpp）	LPIPS（10^{-2}↓）
VVC	0.045	36.900	0.035	58.300
	0.067	29.600	0.040	56.300
	0.075	27.400	0.047	54.000
	0.095	23.100	0.055	51.700
E2E[13]	0.039	33.400	0.016	62.800
	0.067	26.300	0.026	60.200
	0.071	25.600	0.035	53.300
	0.092	24.600	0.052	49.800
LCIC[8]	0.046	27.900	0.036	54.300
	0.055	26.800	0.046	52.000
	0.064	26.100	0.053	50.900
	0.074	25.900	0.061	50.300
SPM[9]	0.049	25.100	0.015	31.000
	0.063	24.100	0.018	29.400
	0.079	23.400	0.025	28.100
	0.110	23.100	0.036	27.800

E2E：端到端可学习图像编码方法
LCIC：分层概念编码方法
LPIPS：可学习感知图像块相似度
SPM：语义先验模型
VVC：多功能视频编码

概念编码方法能够在特定领域以极低的比特率获得更高的视觉重构结果。此外，与FFHQ相比，LCIC在更具挑战性的ADE20K内容上表现一般，而FFHQ包含大量面部语义区域。相比之下，SPM在具有不同语义区域和纹理的挑战性场景中实现了重建质量的显著提高，验证了所提出的语义先验建模机制的有效性。此外，在ADE20K室外测试集的LPIPS指标上，VVC、SPM[8]和最新研究成果CCL[16]的率失真曲线如图5所示。对比结果验证了应用一致性对比学习方法所带来的重建质量的提高。大量实验表明，与之前的工作相比，概念图像编码在高效视觉表征学习、高效图像压缩（<0.1 bpp）、更好的视觉重建质量以及智能视觉应用（如控制和分析）方面都具有一定优势。

2 概念视频压缩

由于深度生成模型的强大功能，许多方法[15]将视频序列映射为潜在表示，并通过生成网络实现低比特率压缩重构。KONUKO等[22]基于FOMM等图像动画模型[23]，开发了一种用于视频会议的生成式压缩框架。WANG等[24]也提出了一种用于视频会议的说话人头部合成模型，通过自适应地从输入视频中提取三维关键点，实现了与H.264/AVC[25]相当的视频质



（a）SPM vs CCL

（b）VVC vs CCL

CCL：基于一致性对比学习的概念编码方法　SPM：基于语义先验建模的概念编码方法
LPIPS：可学习感知图像块相似度　VVC：多功能视频编码

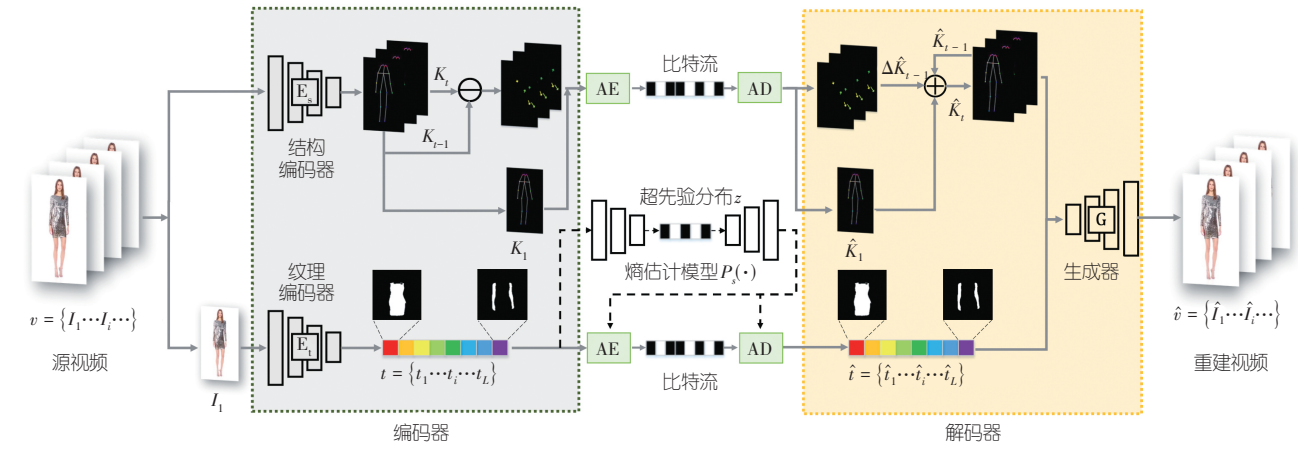
▲图5　SPM[8]、CCL[16]和VVC的率失真曲线

量，而带宽只有后者的十分之一。然而，在极高的压缩比（例如1 000倍）下实现高视觉质量的视频压缩框架仍待研究。

受近年来概念图像压缩的启发，DHVC[9]首次尝试利用分离的视觉表征进行极低码率的人体视频压缩，提出了概念视频压缩方法。在编码端，输入视频序列被分解为结构和纹理表示进行高效压缩。结构表示采用预先训练好的结构编码器来估计每个帧的人体姿势关键点，与传统视频编解码器中的运动矢量类似，计算每个关键点坐标的位移，作为两个帧之间运动信息的特征。为了节省比特率，编码时只传输第一帧的结构表示和后续帧的运动特征。纹理方面，纹理编码器将第一帧提取为语义级纹理表示，用于表示输入视频序列的纹理信息。为了确保所有帧的纹理一致性，引入对比学习[26]来优化纹理表示。在解码端，结构表示通过迭代进行重建，生成器根据纹理表示和结构表示还原视频。最后，引入纹理表示的熵估计模型，与对比学习相结合，建立率失真优化的端到端视频压缩框架，实现更有效的码率节省和更好的重建效果。

如图6所示，人体的主要结构信息可以通过人体姿势关键点有效地表示出来。结构编码器Es采用预先训练好的姿态估计器[27]，提取每帧图像的结构信息作为紧凑的结构表示。纹理编码器Et将图像帧提取为纹理表示。为了更好地捕捉每帧图像的纹理细节，分解组件编码（DCE）模块[28]被用来嵌入语义感知纹理表示。

为了确保同一视频中所有帧的纹理一致性，文献[26]在训练纹理编码器Et时引入了对比学习。选择同一视频中的帧作为正样本，其他视频的帧被视为负样本。此外，该框架



AD：算数解码器　AE：算数编码器

▲图6　概念视频压缩框架

还提出了基于语义层面的对比学习，并利用公式（1）计算语义级信息归一化交叉熵（infoNCE）损失[29]：

$$L_{cst}=-\sum_{i=1}^{L}\log\frac{\exp(t_i\cdot t_i^{-}/\tau)}{\sum_{j=1}^{Q}\exp(t_i\cdot t_{ij}^{-}/\tau)},\tag{1}$$

其中，t_i、t_i^{+}、t_i^{-}、τ、L和Q分别表示输入帧、同一视频中的另一帧、不同视频中的其他帧、温度参数、图像语义区域数量和负样本集的长度。这种技术使编码器既能利用正样本对(t,t^{+})的相似性，又能利用负样本对(t,t^{-})的不相似性。按照MoCo[26]，使用一个队列来存储先前输入帧的负样本t_i^{-}。这样，该模块就能在小批次下高效地进行对比学习。

表2列出了在Fashion数据集和TaichiHD数据集上测试LPIPS和DISTS指标的平均结果。其中，VVC和ArtAni方法压缩的比特率略高于DHVC[9]，但DHVC[9]使用框架仍在超低比特率下达到LPIPS和DISTS分数最低的性能，优于其他压缩框架。此外，表2中的定量结果进一步验证了将对比学习与压缩技术相结合可获得更好的视觉质量。总的来说，由于DHVC[9]将纹理和结构表征分离开来，因此与之前的方法相比，DHVC[9]实现了更高的视觉质量，从而产生了更清晰的结果，保留了更多细节，如面部特征和复杂背景纹理等。

▼表2　视频压缩方法比较（其中，分数越低代表视觉质量越好，w/o c.表示使用的模型不使用对比学习技术）

	Fashion[30]		Taichi[23]	
	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓
VVC	0.268 7	0.300 9	0.314 7	0.270 9
ArtAni	0.177 7	0.227 3	0.301 1	0.249 1
Proposed (w/o c.)	0.110 9	0.168 7	0.215 3	0.220 6
Proposed	0.102 8	0.160 4	0.202 8	0.198 7

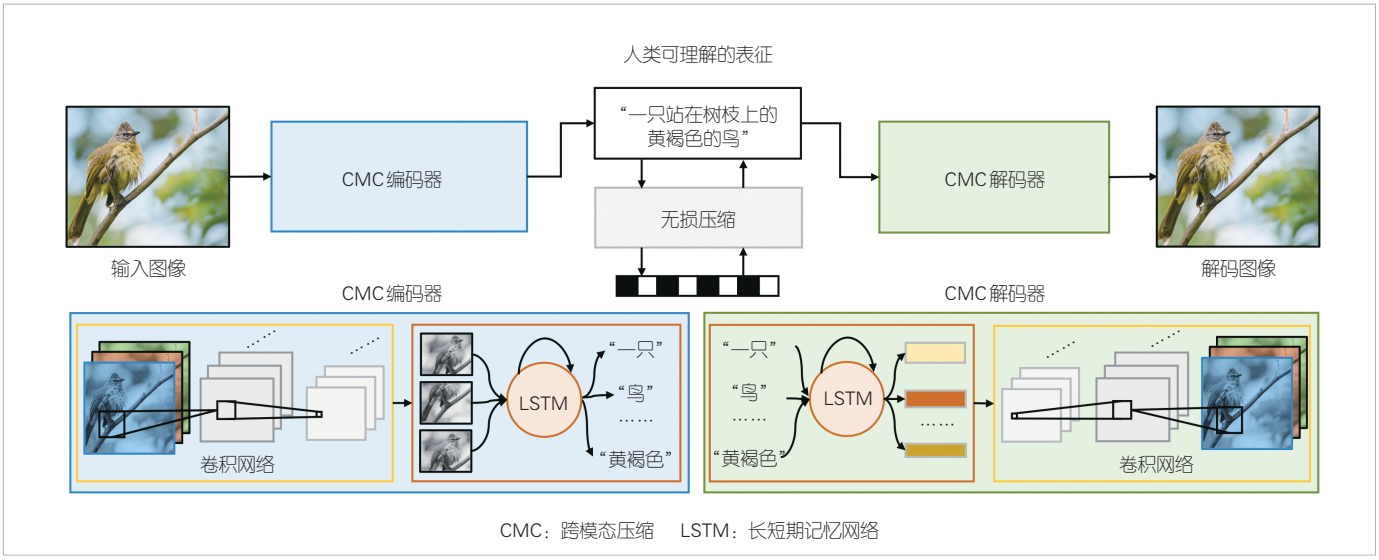
DISTS：深度图像结构与纹理相似度　VVC：多功能视频编码
LPIPS：可学习感知图像块相似度

3 跨模态语义编码

概念压缩框架将视觉内容编码为从深度神经网络中提取的潜在变量等表示形式，这些表示形式对人类来说不易理解。人类可理解的表示形式，如文本、草图、语义图和属性等，在各种应用中具有重要意义，比如语义监测和以人为中心的应用。语义监测旨在监测语义信息，例如身份识别、人流量或车流量，而不是原始信号或潜在变量。以人为中心的应用旨在直接向人类用户传达视觉数据中蕴含的人类可理解信息。因此，我们提出跨模态压缩（CMC）[10]，旨在以超高压缩比将高度冗余的视觉数据转化为紧凑的、人类可理解的表示形式。

如图7所示，CMC框架由4个子模块组成：CMC编码器、CMC解码器、压缩域编码器和压缩域解码器。压缩过程也包括4个步骤：首先，CMC编码器将原始信号压缩成紧凑的、人类可理解的图像表示；其次，压缩域编码器以无损方式将图像表示编码为比特流；然后，压缩域解码器以无损方式从比特流中重建图像表示；最后，CMC解码器以语义一致的方式从图像表示中重建信号。CMC框架通过训练找到一个紧凑的压缩域优化了比特率，并在CMC编码器和解码器中保留语义以减小失真。

随着图像字幕技术[31]和文本引导图像生成技术[32]的发展，从图像生成高质量文本和从文本生成高质量图像变得更加可行。由此，我们建立了一个高效的图像-文本-图像的



▲图7　跨模态压缩框架示意图

CMC范式，将图像压缩到文本域，而文本域具有通用性、普遍性和人类可理解性的特征。具体来说，采用经典的CNN-RNN模型[31]作为CMC编码器将图像压缩为文本。其中，CNN以图像为输入，用于提取图像特征，并将提取到的图像特征输入RNN以自动渐进的方式生成文本。哈夫曼编码[33]可用作组合域压缩的编码器/解码器，以无损方式减少文本的静态冗余。使用文本到图像的生成方面表现出色的AttnGAN[32]作为CMC解码器，用于从文本重建图像。通过在多个数据集上进行大量实验，CMC框架的有效性得到了验证。该模型以超高压缩比（4 000~7 000倍）获得了令人鼓舞的重构结果，显示出比广泛使用的JPEG基线[34]更好的压缩性能。

4 结束语

智能视频编码发展快速，未来或可开发出更先进的模型，进一步提高视觉信号的编码和表示效率。概念编码及跨模态编码技术提高了对使用压缩数据的下游任务的支持能力，开发了具有高度可解释性的潜在表征。通过使用这种表征，未来有可能开发交互式编码技术，从而实现一系列新颖的应用，如内容编辑和身临其境的交互，使其能够提供超越传统视频压缩方法的多功能特性和功能，为压缩领域带来了新的机遇。然而，智能视频编码领域也面临许多新的挑战，包括：1）数据安全性问题。智能视频编码从涉及信号信息的网络中提取的潜在表示可用于重建整个视频流，但这种表征没有加密，存在敏感信息泄露的风险。因此，可信和稳健的编码网络设计在实际应用中发挥着核心作用。2）模型泛化能力问题。考虑智能视频编码的落地实施和部署，关键是确定智能视频编码如何在实际应用领域得到应用，寻找到智能视频编解码器可以满足多样化需求的应用场景。例如，一些为户外场景训练的智能视频编解码器可能并不适合用于编码面部图像。但在实际应用中，采用多个模型进行场景适应是不切实际的。此外，推动智能视频编码标准与其他媒体数据标准相协调，将有助于智能视频编码在实际领域（例如移动设备上的短视频和沉浸式媒体应用）上应用的进程。3）标准制定问题。为了使符合智能视频编码标准的终端和系统能够无歧义地解码潜在表征，有必要通过定义适当的规则并给其分配相应语法元素来实现标准化。在系统层面，结构、语义和文本表征应由兼容的结构、语义或文本解码器正确解析。同时，符合智能视频编码标准的网络应能理解和处理智能模型层面的潜在表征含义。然而，如何可视化或分析高度紧凑的潜在表示比特流，评估现有智能视频编解码器的语义一致性是一项巨大的挑战。一些专业应用可能还需要对潜在表示域进行限制或扩展。如何支持特定领域的扩展和专业化，同时确保无歧义的一致性验证是一个关键问题，需要继续研究和探索。

参考文献

[1] NETRAVALI A N, STULLER J A. Motion-compensated transform coding [EB/OL]. [2023-12-12]. https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1979.tb02277.x

[2] MA S W, ZHANG X F, JIA C M, et al. Image and video compression with neural networks: A review [J]. IEEE transactions on circuits and systems for video technology, 2020, 30(6): 1683 – 1698. DOI: 10.1109/TCSVT.2019.2910119

[3] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets [C]//The 27th International Conference on Neural Information Processing Systems. ACM, 2014: 2672 – 2680

[4] KINGMA D P, WELLING M. Auto-encoding variational Bayes [EB/OL]. (2023-10-18) [2024-05-01]. http://arxiv.org/abs/1312.6114

[5] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models [C]//The 34th International Conference on Neural Information Processing Systems. ACM, 2020: 6840 – 6851

[6] CHANG J H, MAO Q, ZHAO Z H, et al. Layered conceptual image compression via deep semantic synthesis [C]//IEEE International Conference on Image Processing (ICIP). IEEE, 2019: 694 – 698. DOI: 10.1109/ICIP.2019.8803805

[7] CHANG J H, ZHAO Z H, JIA C M, et al. Conceptual compression via deep structure and texture synthesis [J]. IEEE transactions on image processing: a publication of the IEEE Signal Processing Society, 2022, 31: 2809 – 2823. DOI: 10.1109/TIP.2022.3159477

[8] CHANG J H, ZHAO Z H, YANG L B, et al. Thousand to one: semantic prior modeling for conceptual coding [C]//IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2021: 1 – 6. DOI: 10.1109/ICME51207.2021.9428366

[9] WANG R F, MAO Q, WANG S Q, et al. Disentangled visual representations for extreme human body video compression [C]//IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2022: 1 – 6. DOI: 10.1109/ICME52920.2022.9859831

[10] LI J G, JIA C M, ZHANG X F, et al. Cross modal compression: towards human-comprehensible semantic compression [C]//The 29th ACM International Conference on Multimedia. ACM, 2021: 10.1145/3474085.3475558. DOI: 10.1145/3474085.3475558

[11] KRÜGER N, JANSSEN P, KALKAN S, et al. Deep hierarchies in the primate visual cortex: what can we learn for computer vision? [J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8): 1847 – 1871. DOI: 10.1109/TPAMI.2012.272

[12] ZHANG Y Z, HAN K, WORTH R, et al. Connecting concepts in the brain by mapping cortical representations of semantic relations. [EB/OL]. (2023-10-18) [2024-05-01]. https://www.nature.com/articles/s41467-020-15804-w

[13] GREGOR K, BESSE F, JIMENEZ REZENDE D, et al. Towards Conceptual Compression [C]//The 30th International Conference on Neural Information Processing Systems. ACM, 2016: 3556 – 3564

[14] GREGOR K, DANIHELKA I, GRAVES A, et al. DRAW: a recurrent neural network for image generation [C]//The 32nd International Conference on Machine Learning. PMLR, 2015: 1462 – 1471

[15] LOMBARDO S, HAN J, SCHROERS C, et al. Deep generative video compression [C]//The 33rd International Conference on Neural Information Processing Systems. ACM, 2019: 9287 – 9298

[16] CHANG J H, ZHANG J, XU Y M, et al. Consistency-contrast learning for conceptual coding [C]//The 30th ACM International Conference on Multimedia. ACM, 2022: 2681 – 2690
[17] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric [C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 586 – 595. DOI: 10.1109/CVPR.2018.00068
[18] LI Y, WANG S Q, ZHANG X F. Quality assessment of end-to-end learned image compression [C]//The 29th ACM International Conference on Multimedia. ACM, 2021: 4297 – 4305
[19] MINNEN D, BALLÉ J, TODERICI G D. Joint autoregressive and hierarchical priors for learned image compression [C]//The 32nd International Conference on Neural Information Processing Systems. ACM, 2018: 10794 – 10803
[20] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019: 4396 – 4405. DOI: 10.1109/CVPR.2019.00453
[21] ZHOU B L, ZHAO H, PUIG X, et al. Scene parsing through ADE20K dataset [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 5122 – 5130. DOI: 10.1109/CVPR.2017.544
[22] KONUKO G, VALENZISE G, LATHUILIÈRE S. Ultra-low bitrate video conferencing using deep image animation [C]//Proceedings of ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 4210 – 4214. DOI: 10.1109/ICASSP39728.2021.9414731
[23] SIAROHIN A, LATHUILIÈRE S, TULYAKOV S, et al. First order motion model for image animation [C]//The 33rd International Conference on Neural Information Processing Systems. ACM, 2019: 7137 – 7147
[24] WANG T C, MALLYA A, LIU M Y. One-shot free-view neural talking-head synthesis for video conferencing [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021: 10034 – 10044. DOI: 10.1109/CVPR46437.2021.00991
[25] RICHARDSON I. H.264 and MPEG-4 Video compression: video coding for next-generation multimedia [M]. New York, USA: John Wiley & Sons, Ltd, 2003
[26] HE K M, FAN H Q, WU Y X, et al. Momentum contrast for unsupervised visual representation learning [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 9726 – 9735. DOI: 10.1109/CVPR42600.2020.00975
[27] CAO Z, HIDALGO G, SIMON T, et al. OpenPose: realtime multi-person 2D Pose estimation using part affinity fields [J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 43(1): 172 – 182. doi: 10.1109/TPAMI.2019.2929257
[28] MEN Y F, MAO Y M, JIANG Y N, et al. Controllable person image synthesis with attribute-decomposed GAN [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 5083 – 5092. DOI: 10.1109/CVPR42600.2020.00513
[29] OORD A VD, LI Y, VINYALS O. Representation learning with contrastive predictive coding [EB/OL]. (2019-07-22) [2023-10-18]. http://arxiv.org/abs/1807.03748
[30] ZABLOTSKAIA P, SIAROHIN A, ZHAO B, et al. DwNet: dense warp-based network for pose-guided human video generation [EB/OL]. (2019-10-21) [2023-10-18]. http://arxiv. org/abs/1910.09139
[31] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: a neural image caption generator [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015: 3156 – 3164. DOI: 10.1109/CVPR.2015.7298935
[32] XU T, ZHANG P C, HUANG Q Y, et al. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 1316 – 1324. DOI: 10.1109/CVPR.2018.00143
[33] HUFFMAN D A. A method for the construction of minimum-redundancy codes [J]. Resonance, 2006, 11(2): 91 – 99. DOI: 10.1007/BF02837279
[34] WALLACE G K. Overview of the JPEG (ISO/CCITT) still image compression standard [J]. Proceedings of SPIE. The International Society for Optical Engineering, 1990: 1244. DOI: 10.1117/12.19537

作 者 简 介

郭怡琳，北京大学在读硕士研究生；主要研究领域为图像生成与图像压缩。

常建慧，北京大学在读博士研究生；主要研究领域为图像生成与图像压缩；在国际高水平会议及期刊上发表论文10余篇。

黄成，中兴通讯股份有限公司资深系统架构师、新一代视频编码标准预研项目经理，现任CCSA TC1 WG3信源编码组副组长、AVS系统组联合组长、TC28 SC29信标委多媒体分委会委员；主要研究方向为视频视觉编码与智能媒体业务系统。

马思伟，北京大学博雅特聘教授、视频与视觉技术国家工程研究中心副主任、国家杰出青年科学基金获得者、IEEE Fellow，并担任AVS标准视频组长，是“863计划”、国家重点研发计划首席专家；主要研究方向为视频编码与处理。