引用格式：王崇宇，毛琪，金立标．基于生成对抗网络的图像视频编码综述[J]．中国传媒大学学报（自然科学版），2022，29(06)：19-28.
文章编号：1673-4793(2022)06-0019-10

基于生成对抗网络的图像视频编码综述

王崇宇，毛琪*，金立标
（中国传媒大学媒体融合与传播国家重点实验室，北京 100024）

摘要：图像视频编码是多媒体信号处理中重点研究的问题之一，旨在高效、紧凑地表达数据，同时最大程度降低编码失真，节省传输与存储成本。经典的图像视频编码技术自上世纪七十年代起形成基于块的“预测-变换-熵编码”的混合编码框架，每一步均需要人工设计算法分别进行优化，实现像素级别的保真，然而其在低码率下由于量化丢失掉大量高频信息，会产生模糊、块效应等令人无法接受的压缩失真。近年来，基于生成对抗网络的图像视频编码的研究取得了较大的进展。相比经典方法，生成对抗网络在低码率下能够较好地弥补高频纹理细节。本文系统地梳理了基于生成对抗网络的图像视频编码的技术和进展，分别从基于全神经网络的端到端编码、生成对抗网络、基于生成对抗网络的图像视频编码三个方面进行了综述介绍，同时对基于生成对抗网络的图像视频编码的未来发展趋势进行了分析与展望。

关键词：生成对抗网络；图像视频编码；神经网络
中图分类号：TP391　**文献标识码**：A

Review on image and video coding via generative adversarial networks

WANG Chongyu, MAO Qi*, JIN Libiao
(State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China)

Abstract: Image and video coding is a primary research field in multimedia signal processing, whose objective is to efficiently and compactly represent data while reducing coding distortion and reducing transmission and storage costs. Traditional image video coding technology has developed a block-based hybrid "prediction-transform-entropy" coding framework which optimizes each step separately to achieve pixel-level fidelity. Quantization, however, loses a significant amount of high-frequency information at low bit rates, resulting in blurring, block effects, and other unacceptable compression distortions. A significant amount of progress has been made in recent years in the study of generative adversarial networks (GANs) for video and image coding. Compared with classical methods, GANs are able to compensate for high-frequency texture details at low bit rates. In this paper, the authors review the progress made in end-to-end coding using neural networks and GANs, and the techniques and progress associated with image video coding using GANs. Future growth trends are also assessed and forecasted.

Keywords: generative adversarial network; image and video coding; neural networks

---

基金项目：中国传媒大学国家重点实验室专项项目（CUC22GZ035）；国家自然科学青年基金项目（62201522）
作者简介（*为通讯作者）：王崇宇（1999-），男，硕士研究生，主要从事智能信息处理研究。Email：wcy19990623@gmail.com；毛琪（1995-），女，博士，讲师，主要从事视频编码、生成对抗网络的研究。Email：qimao@cuc.edu.cn

1 引言

近年来,以5G为代表的多媒体通信的革新与以深度学习为代表的人工智能技术的发展,催生出以移动终端为支撑的视频分享和通讯平台,其涵盖了视频直播、短视频、社交视频、视频通话、视频会议等众多多媒体应用. 此外智能安防、智慧交通、智慧城市为导向的监控视频、自动驾驶、数字视网膜等新型多媒体应用也开始出现在大众的视野。图像/视频数据正呈现井喷式增长。图像/视频编码是多媒体应用处理的核心技术,旨在高效、紧凑地表达数据,同时最大程度降低编码失真,节省传输与存储成本。自1948年香农建立信息论与编码理论开始,图像视频编码技术便开始蓬勃发展,表1所示。如图1所示,经典图像编码标准JPEG、JPEG2000、BPG和经典视频编码标准H.26X、MPEG、AVS等均基于块的"预测-变换-熵编码"混合编码框架,包括块分割、预测、变换(离散余弦变换(DCT)、离散小波变换(DWT)等)、量化、熵编码、环路滤波等模块,框架中的每个模块均需要通过人工设计。尽管以VVC、AVS3为代表的新一代视频编码标准在性能上与上一代编码标准相比取得了约50%的提升,但由于框架本身的束缚,经过几十年的演进发展,这些模块的设计、实现成本和复杂度越来越高,经典编码框架正面临压缩性能进一步提升的瓶颈。

面向实际的应用场景,例如移动端的视频通话、视频直播、视频会议等,当同时在线大量用户或者网络环境不理想等带宽受限的情况下,经典编码框架在低码率下解码的视频效果主观质量非常差。它们在低码率(比特数<0.1bpp)和极低码率(比特数<0.01bpp)下编码会丢失大量的高频信息,出现量化失真导致的模糊、块效应、颜色失真等难以接受的解码视频。因此,如何提升极低码率下的主观感知质量和编码效率,是目前视频编码应用的难点和瓶颈。随着数据的海量化和计算机视觉技术的发展,越来越多的机器参与到智能处理图像视频信息中。然而,经典编码框架是面向人类视觉的像素级优化,无法很好地支持各式各样的机器视觉需求。

随着人工智能特别是深度学习的兴起与发展,研究人员尝试将神经网络加入到图像视频编码中,利用其数据驱动、机器视觉友好等特点,实现更加智能,更加高效的图像视频编码。当前基于人工智能的图像视频编码主要集中在混合神经网络编码与全神经网络编码的研究。针对全神经网络编码,生成模型特别是生成对抗网络在极低码率下能够较好地弥补高频纹理细节,甚至可以直接利用紧凑的特征生成高感知的图像/视频,为突破极低码率下的编码效率另辟蹊径,开始受到工业界和学术界的关注。

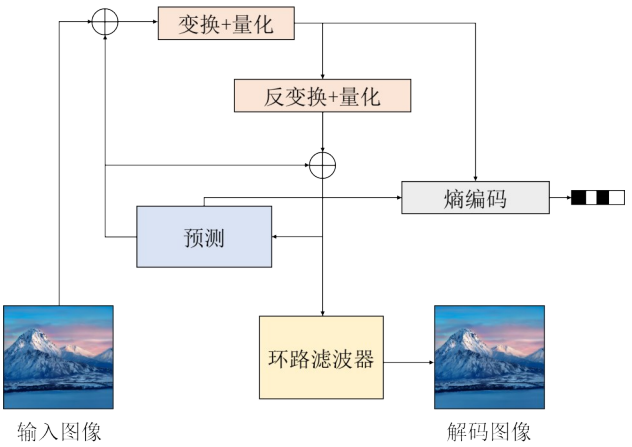

图1 经典基于块的混合图像编码框架

表1 国内外主要视频编码标准发展历程

组织	内容	模型	年份
MPEG	VCD	MPEG-1	1993
	DVD、电视	MPEG-2	1998-2001
ISO、IEC	SD	H.262/13818-2	1994-1998
	HD	H.264/AVC	2003-2008
	4K UKD	H.265/HEVC	2013-2016
	8K、VR/AR	H.266/VVC	2020-
ITU	国际标准	H.120	1984-1988
	视频会议	H.261/H.263+	1990-2000
AVS	国内标准	AVS	2003-2006
	二代标准	AVS2	2012-2015
	超高清、VR	AVS3	2018-2019
AOM	互联网	AV1/AV2	2018

为了更好地梳理基于生成对抗网络的图像视频编码的发展历程,突出其技术重点和难点,探究未来编码工作可能的改进方向。本文对基于生成对抗网络的图像视频编码进行综述。本文第2章对基于全神经网络的端到端编码的研究现状进行分析,第3章概括生成对抗网络的技术发展路线,第4章详细介绍基于生成对抗网络的图像视频编码,最后第5章梳理展望基于生成对抗网络的图像视频编码目前面临的挑战以及未来研究方向。

2 基于全神经网络的端到端编码

近五年来,研究者们尝试将深度学习应用到视频编码中,其研究思路主要分为两类:

(1)混合神经网络编码：通过将经典混合编码框架中的某些模块诸如预测和环路滤波替换成通过离线训练之后的深度学习模块来获得更好的编码性能；

(2)全神经网络编码：如图2所示，探索完全基于神经网络的端到端编码框架。

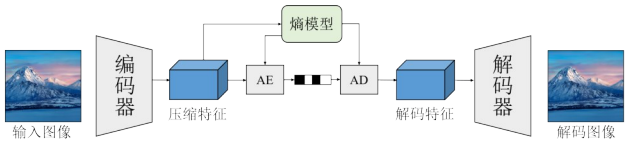

图**2** 基于全神经网络的端到端编码

典型的全神经网络的编码器使用自动编码器，针对香农的率失真(Rate-Distortion)权衡进行端到端的优化，目标是尽可能降低所需的比特率，并提升解码图像的质量。

$$\min R + \lambda D \tag{1}$$

其中，R是由熵估计模型估计出的潜在码的熵，D表示原始图像与压缩图像之间的差异，它们以端到端的方式最小化率失真目标函数。λ表示拉格朗日因子，实现码率和失真的权衡(Trade-off)。自动编码器由编码器E和解码器D构成，编码器将图像x映射到潜在特征$y = E(x)$，解码器用于重建图像$x' = G(y)$。$d(x,x')$是压缩图像产生的失真，一般使用均方误差(MSE)或多尺度结构相似性(MS-SSIM)度量。对y的概率模型$P(y)$使用熵编码算法可以无损地存储其比特流，得到$r(y) = -\log(P(y))$。将E、D和P的参数视为CNN，就可以使用最小化率失真权衡的方式进行训练。当λ设定较小时，图像的压缩比较高，重建图像的感知质量变差：

$$\mathcal{L}_{ED} = \mathbb{E}_{x\sim p_x}\left[\lambda r(y) + d(x,x')\right] \tag{2}$$

早期研究人员发现，由于神经网络模型的训练需要依赖反向传播和随机梯度下降算法，因此损失函数的参数需要处处可微。但量化模块会导致几乎为零的梯度信息，极大影响神经网络模型参数的更新。为了解决这个问题，谷歌研究团队Balle等人[1]首次在训练时引入一种均匀噪声来近似量化误差，使得端到端优化成为可能。模型通过对码率的估计来优化神经网络的率失真函数，提出一种变分逼近压缩框架，通过参数化的概率分布族与交叉熵损失函数，对信息熵的上界进行估计。此外，该团队还提出了广义分歧归一化模块(Generalized Divisive Normalization, GDN)[2]，适合于图像重建问题，可以更好的捕捉图像的统计特性。随后，为了更好地捕获特征图之间的空域冗余，该团队在2018年为其并入超先验建模[3]，这种优先的边信息是经典编码中通用的技术，但是在自编码器中还未得到开发。超先验的网络结构通过与编码器一起端到端优化，实现了网络模型中间码字基于内容的自适应概率估计。该模型也因此成为全神经网络估计码率的通用工具。在此基础上，北大研究团队[6]提出了由粗到细的层次化的超先验建模，以进一步消除空间冗余。然而超先验其感知上下文能力有限，仍然有可能忽略一些相关性。后续研究者们[4,5]还尝试从上下文概率估计模型的方向设计更准确的熵估计模型，并获得了超过HEVC帧内编码的压缩性能。但上下文概率估计模型是根据相邻的m个元素进行概率取值，因此无法建模长期依赖。Cheng等人[7]在2020年使用离散的高斯混合似然来对分布进行参数化，从而消除了特征图中存在未捕获的结构冗余特征，实现了更准确的熵模型，所需的编码位数更少，在网络结构中还采用了注意力模块关注复杂区域以提高性能。

然而，这些研究工作仅局限于独立训练固定码率模型，后续研究者的工作开始思考如何更好地适配不同的码率，提出了一些单一模型下的可变码率方案：例如，Choi等人[8]引入拉格朗日乘数和量化步长作为码率控制参数，通过改变拉格朗日乘数，能够对目标码率进行粗码率配准，而通过调整量化步长，能够更精细化地调整码率。在此基础上，Song等人[9]又提出了一种基于空间特征变换的可变码率压缩框架。

在消除时域冗余方面，研究者们也提出了一些更高效的编码框架。例如：中科大的Chen等人[10]将像素级别的运动估计模块加入到端到端的视频编码框架中。上海交大研究团队的Lu等人[11]利用光流来对运动估计进行建模，并加入了预测残差编码的环节，以进一步去除时域冗余。谷歌研究团队[12]则提出了偏置场(Displacement Field)的概念来代替光流对运动进行建模，并提出基于不同尺度的变换操作来实现视频帧的重建。微软亚洲研究团队[13]提出深度的上下文视频压缩框架，实现从残差编码到条件编码的范式转变。

尽管目前全神经网络视频编码取得了不错进展，但是由于其优化目标主要以基于像素级别的MSE或MS-SSIM作为失真度量，在低码率上得到的重建图像质量主观感知效果较差，此外其码率无法很好地压缩至极低码率(<0.01bpp)。

3 生成对抗网络

生成对抗网络是一种通过对抗性训练学习生成新数据的深度生成模型,由于其无监督、生成质量高的特点,被广泛应用于图像任务中。它由生成器G和鉴别器D两个神经网络组成,如图3所示,生成器通过对抗损失函数不断地从采样信号中生成样本数据,期望得到与真实样本相似的虚假样本,判别器期望区分真实样本和来自生成器的虚假样本。两个神经网络通过对抗的方式进行训练,不断提升自己的生成和鉴别水平,最终达到一个纳什均衡的状态,得到逼真的高分辨率生成图像。

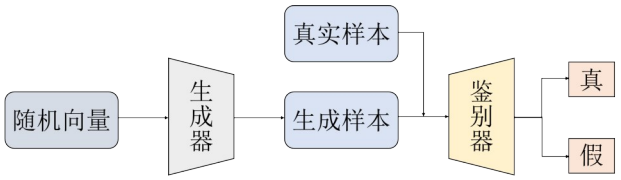

图3 生成对抗网络框架

生成对抗网络这一学习范式最早提出于2014年,Goodfellow[14]使用两个MLP搭建对抗网络,通过KL散度来度量真实数据与生成数据之间的差异。近五年来,对生成对抗网络的研究取得了惊人和长足的发展。表2梳理了生成对抗网络模型的主要研究脉络。

表2 生成对抗网络研究脉络概述

分类	内容	模型	年份
损失函数	KL散度	GAN[14]	2014
	EM距离	WGAN[15]、WGAN-GP[16]	2017
	最小二乘损失	LSGAN[17]	2017
	模式寻找正则项	MSGAN[18]	2019
模型结构	卷积网络	DCGAN[19]	2016
	自编码/变分自编码器	VAE-GAN[20]	2016-
	渐进式网络	ProGAN[21]、StyleGAN[26, 29]	2017-
	自注意力机制	BigGAN[22]、SAGAN[50]	2018-
	结合Transformer	VQ-GAN[25]	2021-
应用发展	图像转换	Pix2Pix[52]、BicycleGAN[53]、MUNIT[54]、DRIT++[55]、CycleGAN[27]、StarGAN-v2[56]、SAVI2I[33]	2016-
	超分辨率	SRGAN[28]	2016-
	基于语义的图像生成	SPADE[57]、SEAN[58]	2019-
	文本到图像生成	StackGAN++[59]、AttnGAN[60]	2018-
	GAN反演	pSp[23]、e4e[24]	2019-

在损失函数方面,研究人员主要解决GAN训练不稳定的问题。WGAN,WGAN-GP[15, 16]使用EM距离,具有优越的平滑特性,在训练初两个数据分布重叠较少时也能够较为准确地刻画彼此之间的距离,改善了梯度消失问题。LSGAN[17]使用均方误差,对距离决策边界较远的图像进行约束,提高了图片的生成质量。MSGAN[18]提出了模式寻找的正则项来缓解GAN模型在训练中的模式坍塌问题,通过最大化生成图像与相应的潜在空间的比率,实现增加生成次要模式样本的机会,提升生成图像的质量和多样性。

在模型结构方面,最具有代表性的模型是StyleGAN系列模型[26, 29],通过可学习的映射网络将高斯噪声分布映射到新的分布,并将其作为可以控制风格的输入,利用自适应实例正则化层(Adaptive Instance Normalization, AdaIN)[30]或解调-调制层加入到生成网络中,以生成更加丰富的纹理细节。最近,研究者们将GAN与Transformer模型结合实现更高分辨率和高逼真度的图像生成:在VQ-VAE模型的基础上,VQ-GAN[25]模型将图像特征利用矢量量化得到序列特征,并利用Transformer对码本的索引进行预测,通过增加感知损失和对抗损失来增强VQ-VAE训练,最终生成逼真的高分辨率图像。

在应用发展层面,最经典的任务是利用GAN进行图像转换(Image-to-Image Translation, I2I),通过替

换目标图像的风格特征并保留源图像的结构特征，能够实现源图像到目标图像域的风格转换。Huang 等人[31]和 Lee 等人[32]首次利用这个思想实现了不同图像域的图像的内容特征和风格/属性特征的分层建模。最近，Mao 等人[33]提出了一种有符号属性向量，能够在不同域的不同映射路径上进行连续转换，实现了跨图像域的连续图像转换。此外，目前 GAN 模型已被广泛应用于图像/视频重建任务包括超分辨率、图像去噪、图像复原、去模糊等以生成逼真的高频纹理信息。Ledig 等人[28]首次提出将生成对抗网络应用于超分辨率任务，使用基于 VGG 的内容损失得到具有丰富纹理效果的生成图像，开启了基于感知质量驱动的图像/视频重建的时代。Mao 等人[34]针对解码图像增强重建，提出基于边缘保持的生成对抗网络，在保持边缘的基础上进一步提升了纹理的丰富性，有效提升了解码图像的人类视觉感知质量。

近年来，利用不同模态的信息进行可控图像生成应用也发展迅速。例如，在基于语义的图像生成方面，Park 等人[57]在 2019年提出空间自适应正则化层，通过语义图生成调制参数来扩展正则化层，有效地在网络中传递语义信息，提升图像生成质量。在此基础上，Zhu 等人[58]进一步提出语义区域自适应正则化层，每个语义区域都可以分别使用一种风格图像作为输入，生成空间上不同的调制参数扩展正则化层，进一步提升生成图像的纹理细粒度。在文本到图像生成方面，Zhang[59]等人使用树状结构排列的生成器与鉴别器堆叠出多级生成对抗网络，以端到端的方式逐级捕获文字的细节信息，实现高质量图像生成。Xu 等人[60]在上述工作的基础上为模型引入自注意力机制，并提出新的图像-文本匹配损失，有效提升了图像不同子区域的细粒度细节。

通过将真实的图像映射到 GAN 预训练模型的隐空间中，GAN 反演（Inversion）任务可以很好地搭建真实图像域与潜在特征空间域之间灵活映射的桥梁，能够直接实现对真实图像编辑。早期的工作直接利用优化的方式来针对每一张真实的图像寻找最佳的隐向量，然而这种方式需要对每一幅图像单独做处理。因此，Richardson 等人[23]提出使用特征金字塔提取出三个层次的语义特征，通过映射网络将特征映射到 W+潜在空间中，然后输入至 StyleGAN 的不同分辨率的合成网络中。这样的架构可以较好地捕捉原始图像的各种细节，使得重建质量有较大提升。

综上，生成对抗网络在学习大量数据先验的情况下既可以生成高频纹理细节，又可以利用不同模态的信息直接生成图像/视频，还可以直接利用预训练 GAN 反演实现真实图像的编辑，在图像视频编码方向展现出非常大的潜力。

4 基于生成对抗网络的图像视频编码

近年来基于生成对抗网络的图像视频编码研究取得了不错的进展。生成对抗网络在图像视频编码任务中主要用来帮助恢复生成图像视频的内容特征、纹理细节、减少块效应。目前研究表明，引入生成对抗网络能够使图像视频重建效果显著，在低码率和极低码率下能得到比经典视频编码标准视觉主观感知更好的解码性能。

基于生成对抗网络的图像视频编码的研究主要有两种典型的思路：第一种是直接利用对抗损失引导优化端到端全神经网络编码，以重建高频纹理细节；第二种是利用生成式驱动实现极低码率的编码，被称为生成式编码，即在编码端将图像表示成更紧凑的特征表示，利用生成模型在解码端直接生成出纹理丰富的图像/视频。下面分别对这两种方法进行介绍和比较。

4.1 作为失真损失项引导端到端编码

目前主流的全神经网络压缩系统使用以像素保真的损失函数（例如均方误差）作为失真度量，缺少对纹理以及全局结构的刻画。这会导致在低码率下，尽管峰值信噪比（PSNR）和 MS-SSIM 这些经典失真度量效果较好，但是图像内容视觉效果比较模糊，主观感知并不理想。如图 4 所示，为了更好地建模纹理细节，研究人员在目标函数中加入对抗损失，利用生成器生成视觉上吸引人的高主观质量的图像。

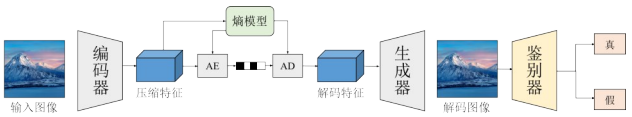

图4 作为失真损失项引导端到端编码

引导编码的失真损失项使用条件生成对抗网络的对抗损失，它对生成器和鉴别器添加边信息 s，学习样本 y 的条件分布 $p_{X|s}$。生成器在边信息 s 的约束下，将样本 y 的分布 p_Y 映射到 $p_{X|s}$。鉴别器输入 (x,s)，判别 (x,s) 是来自 $p_{X|s}$ 还是来自 p_Y。训练目标是让 D 将来自生成器的样本判别为真，损失函数选用非饱和损失函数，能够在训练早期提供较大梯度。

$$\mathcal{L}_G = \mathbb{E}_{y\sim p_Y}\left[-\log\left(D\left(G\left(y,s\right),s\right)\right)\right], \tag{3}$$

$$\mathcal{L}_D = \mathbb{E}_{y\sim p_Y}\left[-\log\left(1-D\left(G\left(y,s\right),s\right)\right)+\mathbb{E}_{x\sim p_{X_0}}\left[-\log\left(D\left(x,s\right)\right)\right]\right]. \tag{4}$$

早期的工作,研究人员主要是在已有端到端编码框架中加入对抗损失作为失真来优化编码,达到在较低码率上通过生成以假乱真的纹理来补充高频信息。Rippel等人[35]首次提出了一种基于GAN的有损图像压缩方法,在一个具有金字塔分析、自适应编码模块和预期码长正则化的自动编码器的基础上,采用启发式对抗训练方式,在低码率下重建具有丰富纹理细节解码图像。Agustsson[36]等人同样提出使用生成压缩(GC)的优化方法代替之前对经典目标(如MS-SSIM和MSE)的优化方法,显著节省了比特流,防止了压缩伪影,在低比特率下取得了令人信服的视觉重建效果。

近两年的工作,研究人员主要侧重于提出新的训练方法以解决引入对抗损失导致的模型训练不稳定的问题。Mentzer[37]等人充分探究了标准化层、生成器和鉴别器架构、训练策略与感知损失。在标准化层方面,使用ChannelNorm代替InstanceNorm缓解生成图片的暗化伪影。在生成器和鉴别器架构方面,将编码器量化后的输出用最邻近上采样进行放大,作为鉴别器的条件生成对抗网络的约束信息,实现了在较低的比特率下高分辨率图像的重建。Lee[38]等人提出一种提高重建感知质量的图像压缩网络训练方法,在引入对抗损失之前,模型首先预训练一个集图像压缩与质量增强为一体的网络结构(EIC-E2E-B),在之后通过对抗性训练获得边缘更清晰,纹理更丰富,更加适应复杂的人类视觉系统感知的图像。与这种两阶段训练方式相似,Iwai[39]等人先通过优化率失真函数来训练编解码器,再单独使用GAN来微调解码器,使得模型训练更加稳定。

4.2 生成式编码

生成式编码利用生成对抗网络模型对紧凑的特征直接在解码端生成出高逼真图像/视频的编码框架,其最初提出的思路是期望在带宽受限的情况下,网络根据信息优先度进行排序并优先存储更高级别的表示,达到用极少的码率来存储信息最紧凑的特征。重建时保持该紧凑特征语义不变的情况下,其余部分直接通过生成模型"想象"填充。

目前学术界正处于"百家争鸣"的态势,不同研究团队提出不同的编码框架。早期的生成式编码工作,主要侧重于编码紧凑性特征,实现极低码率压缩。研究者们对如何进行紧凑表示给出了各自的解决方案。Wu等人[40]利用遮罩器训练网络来指导比特分配,他们将图像压缩后的比特流通过一个卷积神经网络输出重要性矩阵,遮罩通过遮蔽掉非重要区域,督促网络为重要区域分配更多的比特。Santurkar等人[41]将GAN的生成器与变分自编码器的解码器共享参数,通过变分自编码器得到图像到隐空间向量的表示,然后利用GAN合成图像。尽管该工作能够实现极低码率下恢复出满足原始图像语义的图像,然而由于早期的生成模型技术受限,该方法生成图像的感知质量和分辨率非常受限。在上述模型结合变分自动编码器和生成对抗网络的优点的基础上,为了更好地对图像进行紧凑特征的表示,Chang等人[42]首次将图像表示为基于边缘图的结构特征以及低维的纹理特征向量,训练一个端到端的VAE-GAN网络,VAE选用KL损失,GAN选用最小二乘损失,并加入$\mathcal{L}_1$损失和$\mathcal{L}_{Latent}$损失分别约束原始图像、潜码、生成图像三者之间的差距。在编码端,图像被压缩为两层比特流,分别为由变分自动编码器编码出纹理特征的比特流和由边缘图像映射的结构比特流。在解码端,生成器基于纹理特征和重构的边缘图映射直接生成解码图像。在上述工作的基础上,Chang等人[43]提出了一种分层融合GAN(Hierarchical Fusion GAN, HF-GAN)来学习由"粗到精"的学习范式,在重建纹理表示和结构映射后,将纹理层和结构层逐分辨率合成到生成图像中,使得解码的图像生成质量和分辨率都得到进一步提升。此外,为了将码率约束加入到编码框架中进行联合优化,实现端到端的编码。Chang等人[44]利用语义分割图作为结构指导,在每个单独的语义区域内提取基于语义的纹理特征,并利用语义相关性进行更精确的熵估计,实现纹理特征的码率估计。综上,如图5所示,这三个工作均可以视为基于分层特征表示的生成式编码。

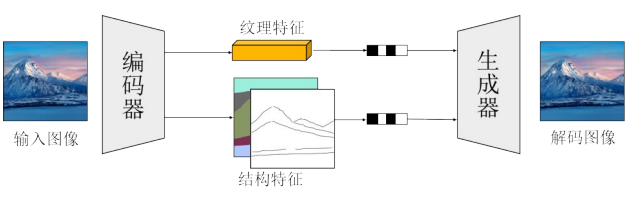

图5 Chang等人提出的基于分层特征表示的生成式编码

在视频方面，英伟达的研究团队[45]首次尝试将基于人脸的生成模型应用于视频会议编码传输中：传输时只传输单一关键帧，并通过提取相应的 3D 人脸关键点、姿态估计以及表情变形估计来建模人脸的姿态、表情的运动，在解码端直接通过生成网络对其他人脸视频帧进行合成。与视频编码标准 H.264 相比，其带宽能够节省 90%。Wang 等人[46]提出在编码端将视频分解为人体关键点结构特征与纹理特征，并利用人体关键点得到运动特征；解码效果如图 6 所示，在解码端，利用生成模型 GAN 实现高质量视频重建。此外该框架利用对比学习监督实现同一视频的相邻视频帧共享单一纹理特征，因此该框架在典型数据集下与最新视频编码标准 VVC 相比，可以实现主观质量较好的极低码率压缩效果。

图 6 英伟达研究团队提出的基于人脸生成编码效果图

由于生成模型本身的特性，可以根据不同的输入信息面向不同的任务需求，生成重建成不同质量的图像/视频。因此，Hu 等人[47]首次提出一种面向人类视觉和机器视觉的可伸缩的图像编码框架，首次尝试将机器视频编码和基于可伸缩特征的图像编码相结合，在人眼视觉质量和机器视觉任务方面都取得优异的效果。将图像表示为边缘和颜色信息，结构特征表示为矢量化的边缘映射，颜色特征表示为从结构特征位置附近采集的稀疏颜色参考像素，利用 GAN 实现对人脸图像的重建，并利用重构的人脸图像进行机器分析。在此基础上，Yang 等人[48]在编码阶段构建更具可伸缩性的颜色特征表示和解码阶段的图像控制效果两个方面对模型加以改进，使得模型可伸缩性进一步增强。如图 7 所示，对于参考像素的选择，通过解码器 D 的反馈确定删除像素的优先级，并根据 SSIM 对像素排序的结果删除彼此距离较远的多个像素。对于极端的训练方式，通过引入对色彩的权衡（Trade-off），在训练时使用遮罩随机遮掉一部分参考像素来模拟不同数量的颜色线索，并在解码端加入 AdaIN 层来影响生成图像的细节与纹理。

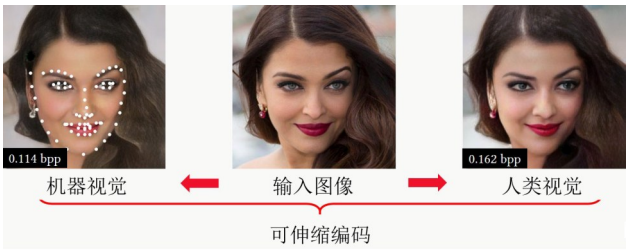

图 7 Yang 等人提出的面向机器视觉和人类视觉的可伸缩编码

此外，Li 等人[49]首次提出跨模态语义压缩（CMC），通过率失真优化将高度冗余的图像或视频转换为一个紧凑的文本描述特征，再使用生成对抗网络从文本域重建图像，由此证明跨模态语义编码的可行性。如图 8 所示，由于文本信息的高效性该方法重建图像只能实现语义上的一致性，与原始图像的主观质量仍具有较大差别。

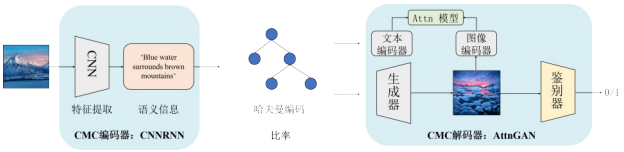

图 8 跨模态语义压缩（CMC）框架

5 总结和展望

随着深度学习和生成对抗网络近几年的快速发展，给图像视频编码注入了新的研究活力。本文首先对基于全神经网络的端到端编码进行介绍，然后梳理了生成对抗网络的主要研究脉络，最后对基于生成对抗网络的图像视频编码方法进行了分析和总结，包括作为失真损失项引导端到端编码和生成式编码。基于生成对抗网络的图像视频编码方法具有广阔的应用前景，包括但不限于极低码率下的视频会议传输系统、极低码率下的直播系统与极低码率下的短视频平台分享等。

目前，生成式编码方法在极低码率下生成高感知的图像视频具有较大的优势，其在编码效率、提取特征可编辑、人机协同、多模态可支持性、质量评价等方面仍具有较大的研究空间。未来，基于生成对抗网络的图像视频编码的探索方向如下：

（1）紧凑性：图像、视频数据携带大量冗余信息，未来编码器在带宽允许的范围内能够选择最高级别、最有价值的信息进行存储，实现表示的紧凑性，提高系统效率。

(2)可扩展性:不同的应用场景对比特率的限制和图像质量的要求各不相同。未来的生成式编码框架能够根据信息的需要灵活地支持各种类型的任务。当比特率约束很紧时,能够强制压缩特征的紧凑性;当带宽充足时,能够忠实地提供高质量重建图像。

(3)多功能性:未来的生成式编码框架能够同时满足机器视觉和人类视觉的双重需求。机器视觉任务具有多样性,编码框架应全面地覆盖不同应用场景和不同用户需求的变化。

(4)泛化性:未来的生成式编码框架不仅仅只局限于特定领域,即使是在语义信息差距大的图像视频数据集中,它也能够保持纹理和语义信息的一致性。泛化性要求生成式编码框架努力实现信息在跨模态意义上的统一。

(5)新的质量评价模型:由于基于生成对抗网络的编码框架的失真与传统编码和基于像素级别优化的压缩失真在视觉感知上具有较大差距,因此未来亟需研究面向生成式视觉特性的有效质量评价模型,由此设计率失真优化方法进一步优化基于生成对抗网络的编解码框架。

参考文献(References):

[1] Ballé J, Laparra V, Simoncelli E P. End-to-end optimized image compression[DB/OL]. arXiv:1611.01704, 2016.

[2] Ballé J, Laparra V, Simoncelli E P. Density modeling of images using a generalized normalization transformation[DB/OL]. arXiv:1511.06281, 2015.

[3] Ballé J, Minnen D, Singh S, et al. Variational image compression with a scale hyperprior [DB/OL]. arXiv: 1802.01436, 2018.

[4] Minnen D, Ballé J, Toderici G D. Joint autoregressive and hierarchical priors for learned image compression [C]// 32nd Conference on Neural Information Processing Systems (NIPS 2018), 2018.

[5] Lee J, Cho S, Beack S K. Context-adaptive entropy model for end-to-end optimized image compression [DB/OL]. arXiv:1809.10452, 2018.

[6] Hu Y, Yang W, Liu J. Coarse-to-fine hyper-prior modeling for learned image compression [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(07): 11013-11020.

[7] Cheng Z, Sun H, Takeuchi M, et al. Learned image compression with discretized gaussian mixture likelihoods and attention modules[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 7939-7948.

[8] Choi Y, El-Khamy M, Lee J. Variable rate deep image compression with a conditional autoencoder [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 3146-3154.

[9] Song M, Choi J, Han B. Variable-rate deep image compression through spatially-adaptive feature transform [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 2380-2389.

[10] Chen Z, He T, Jin X, et al. Learning for video compression [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(2): 566-576.

[11] Lu G, Zhang X, Ouyang W, et al. An end-to-end learning framework for video compression [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43 (10): 3292-3308.

[12] Agustsson E, Minnen D, Johnston N, et al. Scale-space flow for end-to-end optimized video compression[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 8503-8512.

[13] Li J, Li B, Lu Y. Deep contextual video compression[C]// NeurIPS 2021: Proceedings of the 35th International Conference on Neural Information Processing Systems, 2021, 34: 18114-18125.

[14] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [C]//NIPS′14: Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014, 2: 2672-2680.

[15] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks [C]//International Conference on Machine Learning, PMLR, 2017: 214-223.

[16] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans [C]//NIPS′17: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 5769-5779.

[17] Qi G J. Loss-sensitive generative adversarial networks on lipschitz densities [J]. International Journal of Computer Vision, 2020, 128(5): 1118-1140.

[18] Mao Q, Lee H Y, Tseng H Y, et al. Mode seeking generative adversarial networks for diverse image synthesis [C]// Proceedings of the IEEE/CVF Conference on Computer Vision andPattern Recognition, 2019: 1429-1437.

[19] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[DB/OL]. arXiv:1511.06434, 2015.

[20] Pu Y, Gan Z, Henao R, et al. Variational autoencoder for deep learning of images, labels and captions [C]// NIPS′16: Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016: 2360-2368.

[21] Karras T, Aila T, Laine S, et al. Progressive growing of gans for improved quality, stability, and variation [DB/OL]. arXiv:1710.10196, 2017.

[22] Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis [DB/OL]. arXiv:1809.11096, 2018.

[23] Richardson E, Alaluf Y, Patashnik O, et al. Encoding in style: a stylegan encoder for image-to-image translation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 2287-2296.

[24] Tov O, Alaluf Y, Nitzan Y, et al. Designing an encoder for stylegan image manipulation [J]. ACM Transactions on Graphics (TOG), 2021, 40(4): 1-14.

[25] Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 12873-12883.

[26] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 4401-4410.

[27] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]// Proceedings of the IEEE International Conference on Computer Vision, 2017: 2223-2232.

[28] Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4681-4690.

[29] Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of stylegan [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 8110-8119.

[30] Huang X, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization [C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 1501-1510.

[31] Huang X, Liu M Y, Belongie S, et al. Multimodal unsupervised image-to-image translation [C]//Proceedings of the European conference on computer vision (ECCV), 2018: 172-189.

[32] Lee H Y, Tseng H Y, Huang J B, et al. Diverse image-to-image translation via disentangled representations [C]//Proceedings of the European conference on computer vision (ECCV), 2018: 35-51.

[33] Mao Q, Tseng H Y, Lee H Y, et al. Continuous and diverse image-to-image translation via signed attribute vectors [J]. International Journal of Computer Vision, 2022, 130(2): 517-549.

[34] Mao Q, Wang S, Wang S, et al. Enhanced image decoding via edge-preserving generative adversarial networks [C]// IEEE International Conference on Multimedia and Expo (ICME), 2018: 1-6.

[35] Rippel O, Bourdev L. Real-Time Adaptive Image Compression [C]//International Conference on Machine Learning, 2017:2922-2930.

[36] Agustsson E, Tschannen M, Mentzer F, et al. Generative adversarial networks for extreme learned image compression [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 221-231.

[37] Mentzer F, Toderici G D, Tschannen M, et al. High-fidelity generative image compression [J]. Advances in Neural Information Processing Systems, 2020, 33: 11913-11924.

[38] Lee J, Kim D, Kim Y, et al. A training method for image compression networks to improve perceptual quality of reconstructions [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020: 144-145.

[39] Iwai S, Miyazaki T, Sugaya Y, et al. Fidelity-controllable extreme image compression with generative adversarial networks[C]//2020 25th International Conference on Pattern Recognition (ICPR), 2021: 8235-8242.

[40] Wu L, Huang K, Shen H. A gan-based tunable image compression system [C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020: 2334-2342.

[41] Santurkar S, Budden D, Shavit N. Generative Compression [C]//Picture Coding Symposium (PCS), 2018: 258-262

[42] Chang J, Mao Q, Zhao Z, et al. Layered conceptual image compression via deep semantic synthesis [C]//IEEE International Conference on Image Processing (ICIP), 2019: 694-698.

[43] Chang J, Zhao Z, Jia C, et al. Conceptual compression via deep structure and texture synthesis [J]. IEEE Transactions on Image Processing, 2022, 31: 2809-2823.

[44] Chang J, Zhao Z, Yang L, et al. Thousand to one: Semantic

prior modeling for conceptual coding[C]//IEEE International Conference on Multimedia and Expo (ICME), 2021: 1-6.

[45] Wang T C, Mallya A, Liu M Y. One-shot free-view neural talking-head synthesis for video conferencing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 10039-10049.

[46] Wang R, Mao Q, Wang S, Jia C, Wang R, Ma S 2022. Disentangled Visual Representations for Extreme Human Body Video Compression [C]//IEEE International Conference on Multimedia and Expo (ICME), 2022:1-6.

[47] Hu Y, Yang S, Yang W, et al. Towards coding for human and machine vision: A scalable image coding approach [C]//IEEE International Conference on Multimedia and Expo (ICME), 2020: 1-6.

[48] Yang S, Hu Y, Yang W, et al. Towards coding for human and machine vision: Scalable face image coding[J]. IEEE Transactions on Multimedia, 2021, 23: 2957-2971.

[49] Li J, Jia C, Zhang X, et al. Cross Modal Compression: Towards Human-comprehensible Semantic Compression[C]// Proceedings of the 29th ACM International Conference on Multimedia. 2021: 4230-4238.

[50] Zhang H, Goodfellow I, Metaxas D, et al. Self-attention generative adversarial networks[C]//International Conference on Machine Learning PMLR, 2019: 7354-7363.

[51] Mirza M, Osindero S. Conditional generative adversarial nets[DB/OL]. arXiv:1411.1784, 2014.

[52] Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of the IEEE Conference onComputer Vision and Pattern Recognition, 2017: 1125-1134.

[53] Zhu J Y, Zhang R, Pathak D, et al. Toward multimodal image-to-image translation[J]. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 465-476.

[54] Huang X, Liu M Y, Belongie S, et al. Multimodal unsupervised image-to-image translation[C]//Proceedings of the European Conference on Computer Vision (ECCV), 2018: 172-189.

[55] Lee H Y, Tseng H Y, Huang J B, et al. Diverse image-to-image translation via disentangled representations [C]//Proceedings of the European Conference onComputer Vision (ECCV), 2018: 35-51.

[56] Choi Y, Uh Y, Yoo J, et al. Stargan v2: Diverse image synthesis for multiple domains[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 8188-8197.

[57] Park T, Liu M Y, Wang T C, et al. Semantic image synthesis with spatially-adaptive normalization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 2337-2346.

[58] Zhu P, Abdal R, Qin Y, et al. Sean: Image synthesis with semantic region-adaptive normalization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 5104-5113.

[59] Zhang H, Xu T, Li H, et al. Stackgan++: Realistic image synthesis with stacked generative adversarial networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(8): 1947-1962.

[60] Xu T, Zhang P, Huang Q, et al. Attngan: Fine-grained text to image generation with attentional generative adversarial networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 1316-1324.

编辑:王谦