基于生成模型的视频图像重建方法综述

王延文，雷为民，张伟，孟欢，陈新怡，叶文慧，景庆阳
（东北大学计算机科学与工程学院，辽宁 沈阳 110169）

摘　要：基于像素相关性的传统视频压缩技术性能提升空间受限，语义压缩成为视频压缩编码的新方向，视频图像重建是语义压缩编码的关键环节。首先介绍了针对传统编码优化的视频图像重建方法，包括如何利用深度学习提升预测精度和利用超分辨率技术增强重建质量；其次讨论了基于变分自编码器、基于生成对抗网络、基于自回归模型和基于 Transformer 模型的视频图像重建方法，并根据图像的不同语义表征对模型进行分类，对比了各类方法的优缺点及其适用场景；最后总结了现有视频图像重建存在的问题，并进一步展望研究方向。

关键词：视频压缩编码；图像重建；生成对抗网络；变分自编码器；Transformer 模型
中图分类号：TN91
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2022178

Survey on video image reconstruction method based on generative model

WANG Yanwen, LEI Weimin, ZHANG Wei, MENG Huan, CHEN Xinyi, YE Wenhui, JING Qingyang
School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China

Abstract: Traditional video compression technology based on pixel correlation has limited performance improvement space, semantic compression has become the new direction of video compression coding, and video image reconstruction is the key link of semantic compression coding. First, the video image reconstruction methods for traditional coding optimization were introduced, including how to use deep learning to improve prediction accuracy and enhance reconstruction quality with super-resolution techniques. Second, the video image reconstruction methods based on variational auto-encoders, generative adversarial networks, autoregressive models and transformer models were discussed emphatically. Then, the models were classified according to different semantic representations of images. The advantages, disadvantages, and applicable scenarios of various methods were compared. Finally, the existing problems of video image reconstruction were summarized, and the further research directions were prospected.

Keywords: video compression coding, image reconstruction, generative adversarial network, variational auto-encoder, Transformer model

0　引言

随着视频流量的大幅增长，进一步提升视频压缩编码性能的需求十分迫切。传统视频图像的压缩编码算法停留在像素表示阶段，仅针对视频的空间冗余、时间冗余、感知冗余以及信息熵冗余进行处理，无法利用图像的语义信息和感知图像的内容相关性，因此基于像素相关性的编码范式难以进一步提升数据压缩比，进入了技术瓶颈阶段。相较于传统方案，基于生成模型的语义压缩编码方法能够进

收稿日期：2022–07–07；**修回日期**：2022–08–22
通信作者：雷为民，leiweimin@mail.neu.edu.cn
基金项目：中央高校基本科研业务费专项资金资助项目（No.N2216010）；国家重点研发计划基金资助项目（No.2018YFB1702000）
Foundation Items: The Fundamental Research Funds for the Central Universities of China (No.N2216010), The National Key Research and Development Program of China (No.2018YFB1702000)

一步感知视频数据间的统计规律，通过将图像内容转换为低语义冗余的概念表示，如结构、纹理和语义等，利用图像间的结构相似性和先验知识来消除视频图像数据间的语义冗余，从而有望极大提升压缩性能。作为压缩编码的重要环节，视频重建是指解码端根据接收的码流信息恢复出原始视频，是低码率下视频质量的重要保证。目前的视频重建方法可以分为 2 种，一种是基于传统混合编码框架[1-3]的重建方法，利用帧内预测和帧间预测技术结合编码残差来重建视频帧，或者利用超分辨率技术重建图像的高频信息，从而实现质量增强；另一种是基于生成模型[4]和语义分析模型[5-6]的重建方法，根据编码端发送的图像特征描述符，即提取图像的稀疏特征表示或者潜在的特征向量，利用生成模型建立特征空间到像素空间的有效转换，从而实现图像重建。

一般来说，生成模型的目标是根据训练数据学习一个能够模拟该数据集的概率分布，并生成符合该分布的新的样本数据。目前，主流的生成方法有 3 种，一种是基于变分自编码器（VAE, variational auto-encoder）[7]，通过明确的概率估计来拟合真实的样本分布；第二种是基于生成对抗网络（GAN, generative adversarial network）[8]，利用生成器与判别器的相互博弈来训练网络，使其不断逼近真实分布；第三种是基于自回归模型[9]实现图像生成，包括利用卷积来建模像素概率分布和基于 Transformer[10]的网络架构实现图像预测。相对于其他 2 种方案，GAN 不需要对生成分布建立显式表达进而避免复杂的计算[11]。此外，通过语义函数来构建损失函数而非基于像素级的相似度量，能够生成更高质量的视频图像，是目前使用最为广泛的方法。

本文主要针对编码框架中的重建方法进行综述，其中重点介绍生成式的重建方法。首先从传统编码重建方法出发，分析利用深度学习进行优化的预测方法。其次结合几种生成模型，总结其可用于视频图像重建的相关方法。最后通过分析现有的编码重建方法存在的相关问题，讨论进一步的研究方向。

1　基于传统编码框架的视频图像重建方法

传统的视频编码框架是由预测编码和变换编码组成的混合编码框架。其中，预测编码主要包括帧内预测和帧间预测 2 种模式，旨在消除视频数据的空间和时间冗余，变换编码通过对残差数据进行变换量化以消除数据的统计冗余。基于这种混合式编码框架，H.264/AVC（advanced video coding）[1]、H.265/HEVC（high efficiency video coding）[2]、VVC（versatile video coding）[3]编码方案通过探索像素之间的冗余，实现了非常高效的编码效率和良好的重建效果。随着深度学习的不断发展，许多研究者将深度神经网络与传统框架相结合，用于改进其中的某些模块，如帧内预测、帧间预测、环路滤波等，进一步提高编码效率和重建质量。本文主要针对帧内预测、帧间预测和超分辨率重建 3 个方面展开叙述，并在表 1 中总结了基于传统编码框架的视频图像重建的基本原理和主要方法。

1.1　帧内预测

帧内预测旨在根据图像的空间相关性去除空间冗余，利用相邻的重建像素预测当前的编码单元。在传统编码标准中，通过计算率失真代价来优化帧内预测模式，并通过不断精细化划分编码单元以及完善预测模式来增强编码性能。由于传统编码的线性预测模式相对简单，因此对于具有复杂纹理的编码块预测效果不佳。而利用深度神经网络能够进一步提升预测精度，主要包括利用网络优化预测模式，对像素值直接预测以及对传统预测结果的进一步增强。例如，Li 等[12]使用全连接网络直接产生预测像素，并通过训练网络来选择预测模式。Cui 等[13]利用卷积神经网络（CNN, convolutional neural network），以相邻的重建块和 HEVC 的预测单元作为网络输入，对预测结果进一步增强，从而减小预测残差。文献[14-15]等分别利用循环神经网络（RNN, recurrent neural network）和 GAN 增强预测。总体来说，基于

表 1　　**基于传统编码框架的视频图像重建的基本原理和主要方法**

类别	目的	原理	主要方法
帧内预测	消除空间冗余	基于相邻重建像素预测当前编码块	优化预测模式、直接预测像素值、对传统预测结果进行增强
帧间预测	消除时间冗余	基于先前重建帧为参考，利用运动估计和运动补偿完成预测	增强参考帧质量与多样性、双向帧间预测、分数像素插值、光流估计
超分辨率重建	提升图像质量	编码端下采样降低图像分辨率/解码端上采样提高重建图像分辨率	传统下采样+深度网络上采样、深度网络下采样+深度网络上采样

神经网络的方法能够更好地利用编码块的上下文信息，相比于传统编码方法实现了大幅的 BD-rate 增益。

1.2 帧间预测

帧间预测旨在利用视频的时间相关性去除时间冗余，基于运动估计和运动补偿技术实现图像像素值的预测，其主要原理是根据邻近已编码的图像来为当前图像块选择最佳匹配块，并将其作为预测结果。基于神经网络的帧间预测主要通过提升参考帧质量和增强运动补偿来改善编码性能。从改善参考帧的角度来看，除了使用重建帧作为参考帧外，主要通过合成新的参考帧来增加多样性，例如，Zhao 等[16]利用帧速率上转换算法根据重建的双向帧生成虚拟帧作为参考帧；Guo 等[17]提出高低时域的分层编码架构，将低时域的重建帧作为高时域的参考帧。从增强运动补偿的角度来看，Zhao 等[18]进行帧间的双向预测，使用 CNN 非线性方式融合预测块进行双向运动补偿，以提高预测效率；Yan 等[19]利用 CNN 构建分数像素参考网络，由与当前编码帧接近的参考帧生成分数位像素，增强运动矢量估计的准确性。

1.3 超分辨率重建

在低带宽的情况下，可以通过超分辨率技术来保证视频重建质量，具体做法为在编码前对图像进行下采样，然后解码器再上采样到原始分辨率，其整体框架如图 1 所示。早期研究主要通过基于插值、基于字典学习的方式进行超分辨率重建，随着基于深度学习的超分辨率算法[20]的不断成熟，其中一些超分辨率网络被应用于编码框架，相关研究主要集中在解码端的上采样，如 Li 等[21]采用传统滤波方式对图像进行下采样，并设置 2 种模式来决策图像的编码分辨率，然后在解码端利用 CNN 分别对编码块和整个编码帧执行上采样，进一步完善边界处理。Afonso 等[22]通过量化分辨率优化模块来自适应选择输入视频的最佳空间和时间分辨率，使用 VDSR[23]的网络架构重新训练后进行上采样，实现了显著的编码增益。另一种基于超分辨率的编码方案是利用神经网络同时实现上下采样，如 Jiang 等[24]利用 2 个 CNN 协同优化分别实现图像的压缩表示和解码重建，保留更多图像细节。

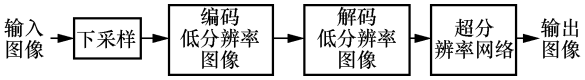

图 1 基于超分辨率的编码重建框架

基于神经网络对传统编码框架进行优化的方法具有很大优势，首先，神经网络能够充分利用视频图像的先验知识，以非线性的学习方式推导数据，优于传统仅依赖于信号处理的模型。其次，神经网络可以更有效地捕获不同处理单元间的相关性，增大时间空间的感受野，从而更好地去除视频的时间空间冗余，提高预测精度。但是在此框架下的编码效率以及重建质量的提升空间有限，无法进一步满足不断增长的用户和市场需求。

目前，基于生成模型的端到端的编码与重建框架，能够以稀疏的图像表示重建原始图像，为探究视频图像语义编码开辟了新的思路，下面，围绕基于生成模型的重建方法展开综述，并在表 2 中对其方法进行分析与比较。

2 基于变分自编码器的视频图像重建方法

变分自编码器（VAE, variational auto-encoder）[7]是一种无监督式学习的生成模型，基于变分贝叶斯推断对输入数据的分布进行建模，其网络架构如图 2

表 2 基于生成模型的重建方法的分析与比较

模型	重建依据	特点
变分自编码器	编解码端学习条件分布，用于拟合真实分布	数学方法明确，易于训练，但对于复杂图像生成样本模糊
生成对抗网络	边缘、颜色、纹理	适用对象更广泛，但目前用于实验的视频分辨率较低
	面部结构特征点、特征域关键点	压缩比更高，但对动作主体要求较为严格，适用场景单一
	语义分割图	同时建立语义与结构表示，但传输语义图会消耗较多码流
自回归像素建模	将图像像素联合分布转换为条件分布，逐像素点预测	善于捕捉图像局部细节，但无法并行计算，重建速度慢且计算成本高
Transformer	直接对像素建模	善于建模图像长期相关性，增大感受野，但难以保证生成图像分辨率
	自回归预测图像的离散视觉标记，将其映射回像素空间	离散数据使图像特征表示更高效，但自回归重建时间相对较长
	掩码视觉 token	更高效地利用数据进行表征学习，但对于视频的应用较少
	利用 Transformer 搭建 GAN 的生成框架	更好地捕捉全局信息，但计算成本较高

所示。在编码部分学习隐变量的分布，首先将输入图像 $\boldsymbol{x}$ 编码为隐变量 $\boldsymbol{z}$，通过学习条件分布 $q(\boldsymbol{z}|\boldsymbol{x})$ 来拟合真实的后验概率分布 $p(\boldsymbol{x}|\boldsymbol{z})$，为方便计算，通常假设 $q(\boldsymbol{z}|\boldsymbol{x})$ 为正态分布，即学习该分布的 2 个参数均值 $\boldsymbol{\mu}$ 和标准差 $\boldsymbol{\delta}$；在解码部分，从隐变量中采样，根据学习到的条件分布 $p(\boldsymbol{x}|\boldsymbol{z})$ 恢复样本数据。

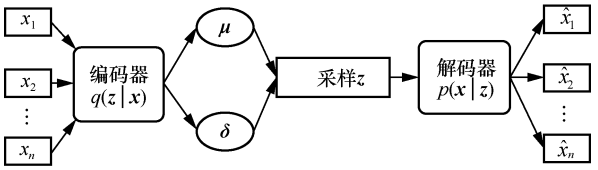

图 2　变分自编码器网络架构

为更有效地控制数据生成，条件变分自编码器[25]通过对编码器和解码器输入 one-hot 向量来表示标签信息，从而实现监督学习，改善重建质量。基于条件 VAE 和条件 U-Net 网络，Esser 等[26]假设图像可由外观和姿态两部分特征来表示，那么图像生成过程可以大致定义为建立关于这 2 个变量的最大后验估计。首先采用 VAE 推断出图像外观，然后利用 U-Net 网络根据外观和姿态信息 2 个分量重建图像。与基于 pix2pix[27]的边缘重建方法相比，该方法能使输出图像与输入图像的边缘保持更高的一致性。

2.1　基于多层 VAE 的重建方式

为更好地近似隐变量的先验和后验概率，一些多层 VAE 模型将隐变量分组为 $\boldsymbol{z}=\{\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_l\}$，同样假设其为高斯分布，逐层自回归建模。因此先验和后验概率可分别表示为

$$p(\boldsymbol{z})=\prod_{i=1}^{l}p\left(\boldsymbol{z}_l\,|\,\boldsymbol{z}_{<l}\right) \tag{1}$$

$$q(\boldsymbol{z}\,|\,\boldsymbol{x})=\prod_{i=1}^{l}q\left(\boldsymbol{z}_l\,|\,\boldsymbol{z}_{<l},\boldsymbol{x}\right) \tag{2}$$

其中，$p(\boldsymbol{z})$ 表示潜在变量 $\boldsymbol{z}$ 的先验分布，$q(\boldsymbol{z}|\boldsymbol{x})$ 表示编码器所学习的近似后验概率。

结合这种分组自回归的推理思想，DRAW[28]采用递归神经网络逐步修正隐变量的分布，其网络结构如图 3 所示。编码端捕获输入图像的显著信息，并采样得到输入的潜在分布，解码器根据接收的条件分布和前一时刻的解码输出，逐步更新生成数据分布。该算法能够生成简单的手写数字，但对于自然图像中的数字生成以及大尺度图像恢复效果有待提升。得益于 DRAW 的生成方法，文献[29]对变分自编码器的潜在特征进行压缩，通过优先存储更高级的抽象表示，实现了图像的概念压缩。

为进一步改善深层 VAE 的参数优化，LVAE（Ladder VAE）[30]设计了一种阶梯网络结构，利用数据之间的依赖性递归修正生成分布，实验结果表明，该网络结构相比于其他自底向上[1]的推理模型更容易优化参数，实现了更准确的对数似然预测和更严格的对数似然界限。在 LVAE 的基础上，BIVA[31]构建了双向推理变分自编码器，通过在生成模型中添加明确的自上而下的路径和在推理模型中添加自下而上的随机推理路径，从而避免了变量崩溃。为进一步提高图像生成质量，NVAE[32]借助文献[33]的统计模型，设计了深度分层的多尺度网络结构，编码器自底向上提取输入表示并自顶向下推断潜在向量，解码器自上而下进行解码，有效捕捉数据的长时相关性。其次提出近似后验残差参数化方法，并

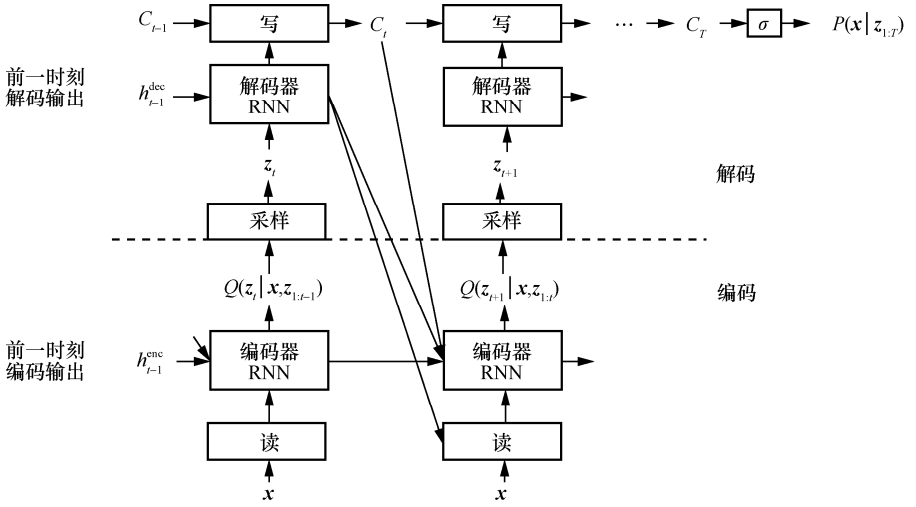

图 3　DRAW 网络结构

为每一层卷积层添加谱正则化保持训练稳定性，首次实现了 VAE 在大的自然图像上的高质量生成。

2.2 基于 codebook 的重建方式

基于 codebook 的重建方式是指为输入图像构建由多个编码潜在向量组成的向量码本，并对其索引实现离散化表示，重建过程即对索引值的预测。VQ-VAE[34]是首个进行离散化表征的 VAE 模型，如图 4 所示，编码器将输入图像编码为潜在表征 $\boldsymbol{Z}_e(\boldsymbol{X})$，同时网络初始构造包含 k 个嵌入向量的编码表，通过共享嵌入空间，利用最近邻查找算法找到与当前潜在变量 $\boldsymbol{z}$ 距离最近的嵌入向量 $\boldsymbol{e}_i$，取其索引值作为当前向量的离散表征，最后经解码器映射回码本中的向量重建图像。这种离散化的数据表示进一步提高了压缩性能和编码效率，为图像重建开创了新范式。同样基于有损压缩的思想，Deepmind 在二代 VQ-VAE[35]中引入层次结构，根据不同大小的潜在空间分别建模图像的局部信息和全局信息，有效提升了图像生成的分辨率。

变分自编码器的重建方法具有更明确的数学理论，可以将数据建模为显式的概率分布，有助于编码器在潜在空间对图像进行压缩表征。但由于 VAE 依靠假设的损失函数和 KL 散度来优化重建图像，当这两项优化失衡时可能会导致后验坍塌，即解码器过于强大，编码器无法提供有效的隐变量表示。此外，对于更复杂的自然图像可能会导致生成样本模糊。

3 基于生成对抗网络的视频图像重建方法

GAN[8]作为一种新的无监督网络框架备受关注。如图 5 所示，GAN 包含 2 个模型，生成器模型 G 与判别器模型 D，生成器根据随机变量生成虚假图片，通过不断学习训练集中真实数据的概率分布，尽可能地提高生成样本与输入图像的相似度；判别器对生成图片与真实图片进行辨别，若输入是真实图片则输出高概率，否则输出低概率，同时将输出反馈给生成器从而指导 G 的训练。二者以博弈的训练方式来分别提升各自性能，使其最终达到纳什均衡，网络损失函数表示为

$$\min_G \max_D V(D,G) = E_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}\left[\log D(\boldsymbol{x})\right] + E_{z\sim P_z(z)}\left[\log\left(1-D\left(G(\boldsymbol{z})\right)\right)\right] \tag{3}$$

其中，$p_{\text{data}}(\boldsymbol{x})$为数据的真实分布，$p_z(\boldsymbol{z})$为输入噪声的向量分布，$G(\boldsymbol{z})$为生成器根据噪声 $\boldsymbol{z}$ 生成的假样本。式(3)中第一项表示判别器识别数据为真实数据，第二项表示判别器识别出数据为生成器生成的虚假数据。当优化判别器 D 时，需固定生成器，使真实数据的判别概率趋近于 1，生成图片的判别概率趋近于 0，因此对应最大化式(3)；当优化生成器 G 时，需固定训练好的判别器，使生成样本接近于真实样本，因此对应最小化第二项。

由于 GAN 生成图像过于随机，缺乏一定限制，无法准确反映训练数据的分布变化，为解决该问题，条件 GAN[36]通过对生成器和鉴别器添加约束条件从而有效指导数据生成，其中条件信息可以是类标签、文本等多模态数据，其损失函数如式(4)所示，y 表示输入条件。相比原始 GAN，条件 GAN 输出更可控，因此更适用于视频图像重建任务。

$$\min_G \max_D V(D,G) = E_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}\left[\log D(\boldsymbol{x}\mid\boldsymbol{y})\right] + E_{z\sim p_z(z)}\left[\log\left(1-D\left(G(\boldsymbol{z}\mid\boldsymbol{y})\right)\right)\right] \tag{4}$$

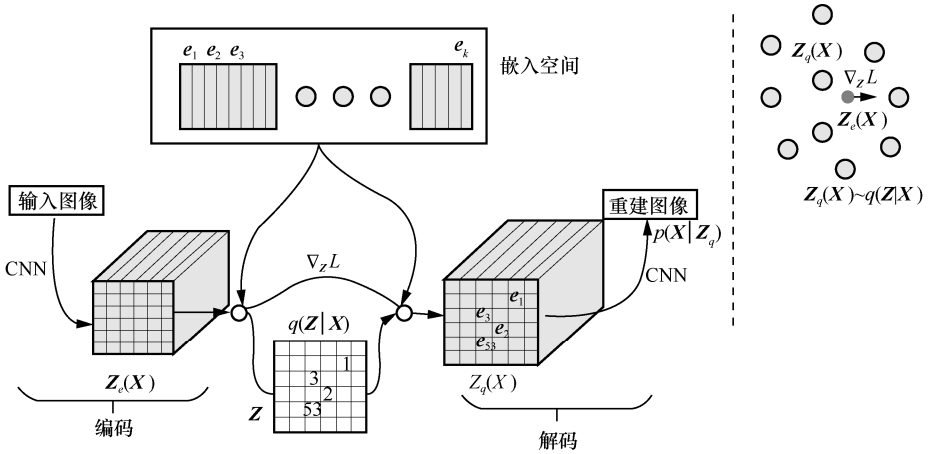

图 4 VQ-VAE 示意

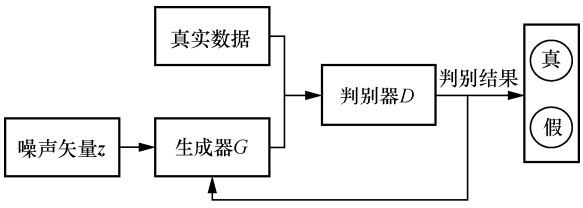

图 5　生成对抗网络示意

得益于 GAN 强大的生成能力，一些基于 GAN 改进的生成方法层出不穷，成为当前视频图像生成效果最为突出的主流方法。例如，pix2pix[27]是最具有代表性的基于 GAN 的图像生成方案之一，文中提出了一个图像转换的统一框架，该框架以输入图像作为条件，利用条件 GAN 生成对应图像。生成器采用了“U-Net”的网络架构，加入残差连接更有效地传递信息；判别器为“PatchGAN”，将图像分块判别，有效建模高频信息。Pix2pixHD[37]在此基础上做进一步改进，通过嵌入多级生成器提升生成图像的分辨率，并采用 3 个作用于不同图像尺度的判别器，分别捕获图像的更大感受野以及精细细节。得益于强大的图像生成能力，这 2 种网络被广泛应用在图像解码重建中，实现了高质量的重建效果。

目前，基于 GAN 的编码重建框架为在编码端提取表征图像语义特征的辅助信息，实现语义压缩，发送方只需传输少量的关键帧和辅助信息；解码端根据辅助信息，使用相关基于 GAN 改进的生成模型恢复图像。根据辅助信息不同，可分为基于边缘、关键点特征以及语义分割图的视频图像重建方法，下面针对 3 种重建方法分别展开讨论。

3.1　基于边缘的视频图像重建

基于边缘的重建借鉴了一部分图像分层的概念，认为结构和纹理是图像中 2 个重要的组成部分，而最常见的几何结构就是边缘。所以一般来说从视觉上可以将图像分为两层：边缘和纹理。按照这种划分思想，就产生了基于图像边缘的重建方式[38-41]。

Hu 等[38]以在编码端提取的边缘和色彩为依据，利用 pix2pix 网络[27]在解码端将二者映射回原始的像素进行图像重建，具体如图 6 所示。在边缘特征提取上，采用基于结构化森林的快速边缘检测[42]来检测边缘的映射，将边缘映射进行二值化，继而将二值化边缘图转换为矢量化表示，从而利用生成模型根据矢量化的边缘图进行重建。该方法在保持高压缩比的同时能够支持机器和人类视觉任务。Kim 等[39]同样以边缘为重建依据，实现了视频重建，不同的是该研究采用“软边缘”，即边缘检测器提取的是带有颜色信息的多级边缘图，而非二进制边缘图。

结合结构和纹理的分层压缩重建方案能够进一步提高图像的保真度。例如，Chang 等[40-41]将视觉数据表示为边缘结构和纹理信息，结合 VAE 和 GAN 这 2 个生成模型实现图像重建。在编码端，利用边缘检测（HED, holistically-nested edge detection）[43]来提取保留图像主要结构信息的边缘图，借助 VAE 提取图像中纹理信息的潜在语义编码。对

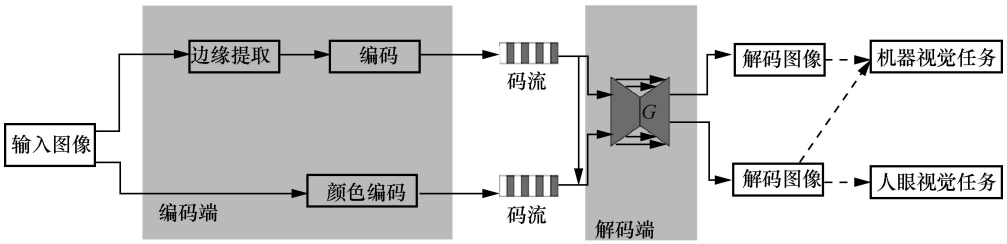

图 6　基于边缘的编码重建框架

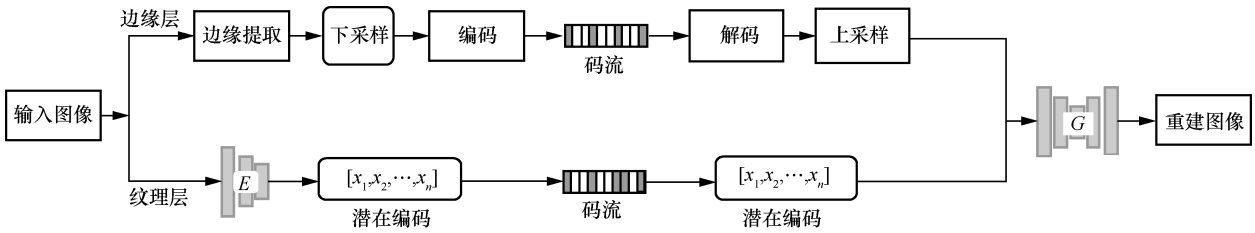

图 7　层间感知的图像压缩和重建网络架构

于重建部分，文献[40]使用最小二乘 GAN 结合获得的低维纹理信息和上采样的边缘图来合成原始图像，其整体网络架构如图 7 所示。文献[41]设计了一个分层融合的 GAN，以残差块为基本单元，通过跳跃连接和分层融合技术逐步提高合成特征图的分辨率。这样的重建方式拥有更好的感知质量，并且保留了原始图像的大部分纹理信息。

用边缘作为紧凑的视觉表征进行编码重建，能够建立视频对象的长时相关性，对图像内容具有更灵活的控制，可极大降低码流。而基于边缘信息的重建方法适用的处理对象和处理任务也更加广泛，包括自然图像合成、人物图像合成等。由于边缘提取效果是保证重建质量的关键，因此对边缘提取算法具有较高要求。目前，以边缘为辅助信息的方法主要集中在图像的压缩重建，对于视频场景，实验的视频分辨率较低，距离实际 1080P、4K 等视频还有很大差距，其次重建的视频帧可能伴随闪烁效应，因此在消除视频时间冗余的同时还需考虑重建视频的完整性与连贯性。尽管目前的视频重建质量仍有待提高，但也为未来视频编码技术提供了新的编码框架。

3.2 基于关键点特征的视频图像重建

关键点特征作为一种常用的人脸结构以及人体姿态表示方法，具有高度抽象且稀疏的特点，尽管缺乏颜色和纹理信息，但能够描述人物的关键结构，也可表示特征域的运动信息，用于辅助视频图像重建，对视频图像压缩编码具有重要意义。使用关键点作为辅助信息的重建方法根据其驱动方式的不同可分为 2 种：一种是使用人脸关键点作为驱动信息，即在面部五官周围设置特定的参考点，使用面部重演技术[44-46]重建人脸图像；另一种是使用关键点表示主体的运动信息，从驱动视频中提取运动特征，利用 talking-head 任务、图像动画、动作迁移等技术重建人物图像。

3.2.1 *以人脸特征点为驱动的重建方法*

以人脸特征点为驱动的重建方法是指通过面部特征点表示运动信息，利用生成对抗网络结合关键帧以及面部关键点进行重建，从而大幅降低视频通话带宽，其网络架构如图 8 所示。Feng 等[47]基于面部重演 FSGAN[48]架构实现重建，选取 1～10 张图像为关键帧传递人脸整体外貌特征和背景特征，提取其他帧的面部关键点用于改变人脸的姿态和表情，并对非关键帧区分出敏感区域加强重建质量。为进一步节省码流，考虑视频内容的长时相关性，可将关键帧编码的码流上传云端或者提前保存本地从而节约实时的传输码流。该方法实现了 1 kbit/s 的良好性能，相比 VVC 编码，可节省 75%码率。文献[49]通过传输扭曲面部分割图，并利用 SPADE[50]改善面部重要区域，实现了移动端低带宽的视频通话。进一步地，Hong 等[51]利用深度图来辅助人脸关键点检测，并学习跨模态注意力指导运动场的学习，使之生成更自然的视频。

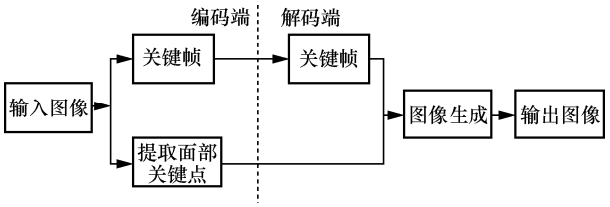

图 8 基于面部关键点的视频重建网络架构

3.2.2 *以视频为驱动的重建方法*

以视频为驱动的重建方法是指将视频分为源视频与驱动视频，分别提供人物的身份信息与运动信息，根据运动信息驱动源视频实现重建。Monkey-Net[52]第一个以自监督方式预测关键点来建模姿态信息，在此基础上，FOMM（first order motion model）[53]根据相邻关键点的局部仿射变换来表征物体运动，并对遮挡部分进行建模实现视频重建。其基本思想都是通过少量的关键点表征不同视频帧之间的运动信息，例如，Wang 等[54]利用 talking-head 模型实现重建，不同于之前的人脸关键点特征表示方式，该模型所提取的关键点是以三维空间分布的形式表征人脸的姿势与表情。整体框架如图 9 所示，首先提取源图像的外观特征，然后通过一阶近似计算驱动关键点相对于源图像关键点的光流，组合多个光流产生最终的光流场用于扭曲三维源特征，最后将扭曲后的特征送入生成器重建图像。由于编码端只需传输关键点，因此在很大程度上节约了传输码流，相比商业 H.264 标准，该方法可以节约 90%的带宽。基于类似思想，Konuko 等[55]同样根据关键点运动信息扭曲参考帧实现重建，并提出了自适应选取参考帧方案，避免由于其他帧与参考帧的时间距离太远导致相关性降低，从而影响重建质量。相较于 HEVC 方案，该方案能够节约 80%的码率。

此外，Few-vid2vid[56]突破了单纯的人脸重建，

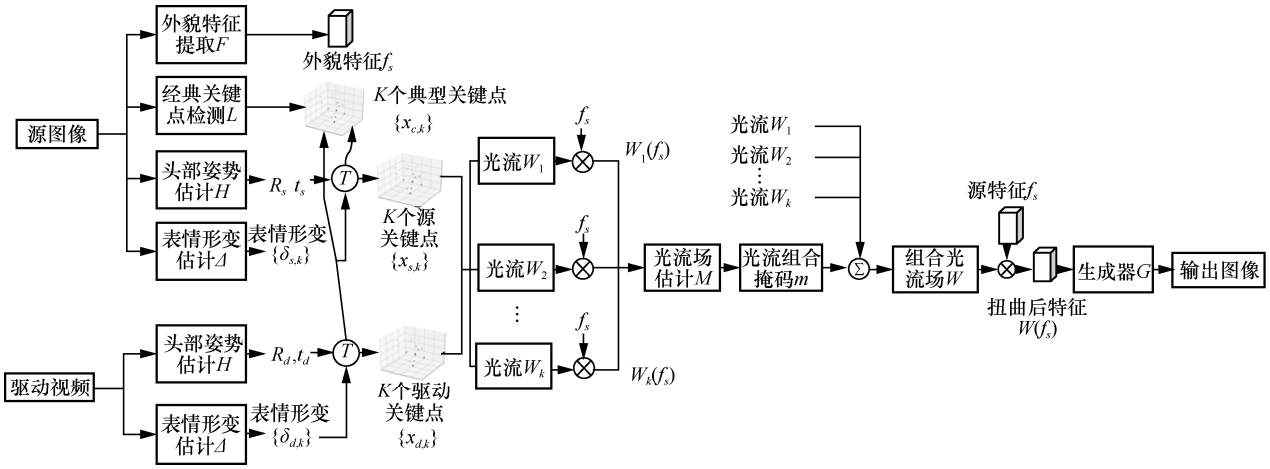

图 9　基于关键点的 talking-head 视频合成整体框架

实现了人体姿态、talking-head 等高质量视频合成。Chan 等[57]完成了 2 个不同人物视频的动作迁移。Xia 等[58]通过学习关键点的稀疏运动轨迹进行重建，实现了一种可伸缩的联合压缩方法。文献[59]通过传输人体姿态和人脸网格信息，利用基于骨骼的动画系统实现重建，最终以动画木偶的形式显示在接收端。Wu 等[60]在重建方法上做出了改进，根据 CovLSTM[61]对（GoP, group of picture）内部的帧间时空相关性进行建模，通过回忆注意力建立特征与关键点之间联系，并将注意力模块的输出作为重建视频的输入条件，基于 pix2pixHD[37]网络来实现重建。但由于该网络对视频序列循环提取抽象特征，适用于非实时视频压缩场景。

相比于使用边缘作为描述图像的低级语义特征，关键点特征能够实现更高的压缩比和更低的传输码流，但由于关键点只表征了位置和方向，无法表征更多的语义信息，因此对动作主体要求比较严格，同时适用的视频场景也相对单一，如只能用于以人物为主体的视频，对于人物姿势变化较大以及背景复杂的视频重建效果欠佳。在网络拥塞导致带宽极低的情况下，借助关键点重建视频的方法对构建高质量实时视频会议、移动端实时视频通话以及流媒体直播具有重要意义，能够进一步节约网络传输资源。

3.3　基于语义分割图的视频图像重建

语义分割作为图像分析的关键步骤，是指对图像中所有像素进行分类，并将同一类别像素用相同颜色表示，从而形成语义分割图，因此语义分割图在一定程度上建立了图像的语义和结构表示，通常也作为视频图像生成的一种辅助条件。例如，Vid2vid[62]根据语义分割图组成的视频来生成视频，将视频到视频合成问题转换为分布匹配问题，通过训练学习使生成视频的条件分布尽可能地与真实视频相接近，以历史图片和语义分割图作为生成器输入合成高清图片。该文实现了合成约 30 s 的 2K 街景视频的超高水平，并且涵盖了视频生成的大部分应用场景。Pan 等[63]采用分治策略实现了基于单一语义标签的视频生成。在语义图像合成中，由于在生成网络中使用归一化层，所以直接将语义分割图送入网络处理会使语义标签激活后变为零，导致语义消失。为解决此问题，Park 等[50]提出了空间自适应归一化，通过自适应学习的参数来调节激活值，保证语义信息的有效性。在此基础上，Zhu 等[64]提出了语义区域自适应归一化，为每个语义区域创建归一化参数，实现对每个语义区域样式的单独控制，进一步提升合成的图像质量和对细节的控制。

基于语义分割图的视频图像合成的应用，文献[65]提出了语义压缩框架，利用 GAN 技术结合压缩的图像表示和语义图重建图像，实现了优于传统图像压缩方法的重建质量，但由于语义图无损压缩进行传输，无疑增加了额外的传输码流。针对此问题，EDMS（encoder-decoder matched semantic segmentation）[66]在编码端与解码端分别进行语义分割，只传输语义重建图像与原图像的残差和图像压缩表示的下采样版本，解码端重新得到语义分割图，并结合残差重建图像，在保证重建图像质量的同时避免了传输语义图耗费码流。虽然这 2 种方法均以语义图为引导重建图像，但主要数据处理还是面向信号级别。为实现面向高层语义的分析处理，Chang 等[67]提出了一种新的对语义先验建模的超低比特率的图像压缩编码方法，如图 10 所示，将输入图

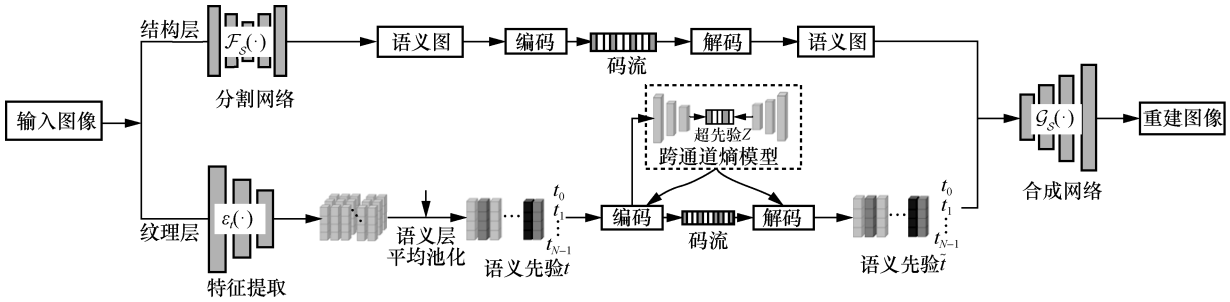

图 10　基于语义先验建模的图像压缩和重建架构

像分为结构层和纹理层 2 种视觉特征，结构层用语义分割图表示，纹理层经过卷积得到高级特征表示，在语义图的指导下，为每个语义区域聚合相应潜在向量作为语义先验，并通过跨通道熵模型建立向量的内部依赖关系，解码端以语义图作为条件，利用生成对抗模型建立语义图与先验之间的分布映射来重建图像，采用感知损失和特征匹配损失保证视觉重建质量，实现了 0.02～0.03 bpp 极低比特率下的感知重建。

本节介绍的利用语义分割图的视频生成模型达到了目前先进的视频合成水平，且涵盖应用场景广泛，包括人物姿态转换、视频风格迁移、视频预测、视频语义属性编辑等。语义分割图建立了每个像素的类别表示，在语义概念层面对图像进行分析，可以进一步增强图像重建质量，适用的场景更为广泛。但相比于之前的边缘和关键点作为重建辅助信息，传输语义图会消耗更多码流。

4 基于自回归像素建模的视频图像重建方法

从概率建模的角度看待视频图像重建，即假设图像 $\boldsymbol{x}$ 由 n 个像素点随机组合形成，那么整幅图像的预测概率可分解为各像素点的预测概率，假设各像素预测概率之间相互独立，则图像 $\boldsymbol{x}$ 预测概率可表示为

$$p(\boldsymbol{x}) = p(x_1)p(x_2 \mid x_1)\cdots p(x_n \mid x_1,\cdots,x_{n-1}) \quad (5)$$

其中，$p(\boldsymbol{x})$ 代表图像 $\boldsymbol{x}$ 的概率分布，符号右边表示预测各像素点的条件概率，重建图像时需按一定顺序逐像素生成。如图 11 所示，每一个像素点的预测都取决于所有之前的像素点，当预测第 i 个像素概率时，则需以前 i−1 个像素作为输入条件。

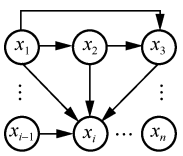

图 11　像素概率预测示意

为提高图像预测准确性，一些研究[68-69]使用深度学习对像素条件概率进行建模，并以最小化图像似然作为损失函数来指导网络训练，其计算式为

$$\min\left(-\ln p(\boldsymbol{x})\right) = -\sum_{i}^{n} \ln p(x_i \mid x_1,\cdots,x_{i-1}) \quad (6)$$

Deepmind[68] 提出了像素循环神经网络（PixelRNN, pixel recurrent neural network）来实现像素概率建模，其中包括采用 2 种长短期记忆（LSTM, long short-term memory）模型来学习图像分布——行 LSTM 和对角 LSTM，前者以一维卷积形式预测该行的像素，而后者以对角线方式扫描图像从而捕捉更多相关信息。但由于 LSTM 运行速度缓慢，导致预测速度减慢。文中的另一个网络 PixelCNN 利用卷积神经网络来建模各像素间的关系，分别沿 2 个方向维度生成像素，并采用特殊的掩码卷积来保证推理顺序。该方法在训练速度上有所提升，但由于利用像素信息有限，生成效果不理想。

除了直接对像素建模实现预测之外，还可以通过先验信息来指导图像生成，如文献[69]中的门控 PixelCNN。

$$p(\boldsymbol{x} \mid \boldsymbol{h}) = \prod_{i=1}^{n^2} p(x_i \mid x_1,\cdots,x_{i-1},\boldsymbol{h}) \quad (7)$$

其中，$\boldsymbol{h}$ 为先验信息，如生成图像的种类、图像高维表征等。

原始的自回归方式是逐通道预测的，忽略了像素之间的相关性。换言之，其将像素预测作为 256 个分类问题，即使预测为相邻的像素也会导致非常大的损失。为解决此问题，PixelCNN++[70]提出了离散逻辑混合似然法，而非基于 256 个通道的回归预测，并通过下采样减少计算量，引入残差连接缩短训练时长。此外，PixelSNAIL[71]将自注意力与因果卷积二者相结合，从而增强对远距离数据的有效建模。

为进一步将自回归模型扩展到大图像，文献[72]提出了子尺度像素网络（SPN, subscale pixel network）进行数据变换，包括图像大小和深度的尺度处理。（VPN, video pixel network）[73]进一步将像素建模的思想应用于视频编码和重建中。将视频表示为关于时间维度、空间坐标以及通道信息的四维张量，同样根据链式法则来预测像素值，为视频生成任务提供了一种通用方案。

此种自回归像素建模的方法在捕捉图像局部细节信息具有一定优势，但仍存在一些不足，主要表现在三方面：其一，由于当前的预测总是取决于之前的信息，因此会导致误差累积；其二，上下文信息过大，需要更为有效的存储和编码方案；其三，对于图像冗余处理仍停留在像素层面，且生成过程需按照固定顺序逐像素生成，无法并行计算，对于处理视频图像等高维数据，速度较慢且计算成本较高。针对这些固有缺陷，可以将其与其他模型结合进行改进，如此前介绍的VQ-VAE[34]利用先降维量化再自回归的方案来减少数据量，以及利用 Transfomer 增强自回归的全局感知。

5　基于 Transformer 的视频图像重建方法

Transformer[10]是一个基于自注意力机制的学习模型，最早用于自然语言处理中。整体保持编码器和解码器的结构，其中编码器由 6 个相同的模块组成，每个模块包含多头自注意力和前馈神经网络 2 个子层，并在子层之间加入残差连接以及层归一化；解码器采取类似的结构，不同的是增加了掩码多层注意力，用于掩蔽未预测的信号。编码器根据一系列注意层获取输入上下文的语义表示；解码器基于前一时刻的解码输出以及编码表征生成输出序列。在之前的重建任务中，主要以卷积的方式实现图像特征提取和重建。相比于卷积的局部感知，Transformer 具有更强的全局感知能力和通用的建模能力。相比于 RNN，Transformer 具有更高效的并行计算，自 2017 年被提出就在视频图像生成中取得了显著效果。

5.1　基于两阶段的重建方式

基于 Transformer 的视频图像生成方法采取与自然语言处理类似的思想，以序列的形式处理图像。Parmar 等[74]首次将 Transformer 应用于图像生成任务，该模型将图像的联合分布转换为像素的条件分布。在编码端，将像素强度表示为 256 个 d 维向量；在解码端，使用局部自注意力建模之前像素与当前像素之间的关系实现各像素点的生成。相比于 PixelCNN[68]，该模型具有对图像长期关系建模和增大感受野的优点。由于此种对像素直接建模难以保证生成图像的分辨率，因此大部分方法还是基于 VQ-VAE[34]构建的两阶段生成方式：第一阶段将图像特征映射为离散标记，第二阶段采用自回归的方式预测标记，将其映射回像素空间。目前，基于 Transformer 的重建方法主要不同表现在对第二阶段的处理方式上，其中 VQGAN[75]将 CNN 与 Transformer 相结合，CNN 用于学习 codebook，Transformer 用于自回归建模，并引入基于块的判别器，利用对抗训练方式保证对于图像局部质量的捕捉，可生成高达百万级像素图像。受自然语言中无监督表征学习的启发，文献[76]证明了 GPT（generative pretraining）模型在图像生成任务中的有效性。VideoGPT[77]通过 3D 卷积和轴向注意力学习视频的离散表征，然后将 GPT 的架构应用于视频的自回归建模。LVT[78]将视频划分为多个不重叠的切片，按照光栅扫描顺序实现自回归预测。文献[79]提出的多模态预训练模型，利用 Transformer 编解码框架为语言、图像和视频定义了一个统一的三维表征，实现了文本到图像、文本到视频以及视频预测等多种视觉合成任务。但这些基于量化的生成模型通常会导致较长的离散序列，为在保证图像率失真性能的同时减小自回归成本，RQ-VAE[80]提出了残差量化的思想，不同于 VQ-VAE[34]的可变大小码本，RQ-VAE 使用固定大小码本，以残差的方式逐渐逼近特征图，解码端使用 Transformer 分别对空间和深度信息进行回归。实验结果表明，在生成高分辨率图像上比之前的自回归模型计算更有效。

5.2　基于掩码建模的重建方式

为了缓解对训练数据的依赖，Bao 等[81]将掩码思想引入图像处理中，基于离散视觉标记重建图像。随后，He 等[82]提出的（MAE, masked autoencoder）证明了掩码在图像表征学习上的有效性，首先对输入图像块随机采样并掩码其余图像块，编码器仅编码未掩码的图像块，然后解码器根据编码的潜在表示以及掩码标记对缺失像素进行重建，其较高的掩码率消除了图像的大部分冗余，从而减少了编码参数。Xie 等[83]提出的 SimMIM 同样使用掩码图像建模来进行自监督学习，与文献[82]

不同的是，SimMIM 编码所有的标记不是仅编码未掩码的部分，解码端使用线性层预测像素值。实验结果表明，仅重建掩码区域可获得更高重建质量，且掩码图像块越小对应的重建质量越高。结合掩码的建模思想和两阶段重建架构，MaskGIT[84]提出了一种双向 Transformer 的图像合成新范式，利用双向自注意力从多个方向生成标记，且掩码部分标记用于下一步的迭代预测，直至生成所有标记。此种双向生成和并行解码的方式极大地提升了回归速度，相比 VQGAN[75]加速了 30～64 倍，同时证明了这种掩码方式在图像重建的有效性，仅需较少标记即可重建出图像的整体信息。由此可见基于掩码的图像建模方式能够高效地利用数据，对于图像表征学习与图像重建具有重要意义，同时选择合适的掩码率有助于节约模型的训练时间与内存消耗。未来，可将其用于视频图像的语义编码中以进一步降低码率。

5.3 基于 Transformer 构建 GAN 的重建方式

上述模型都是以自回归的形式重建图像，这意味着在提高重建时间上有所限制。最近的一些工作[85-88]尝试将 Transformer 与 GAN 相结合，其中文献[85]首次仅利用 Transformer 构建 GAN 实现图像生成。生成器由多个 Transformer 块组成，用以渐进式地提高生成图像分辨率，并通过级联不同大小的图像块实现多尺度鉴别，以防细节信息丢失，但无法生成高分辨率图像，原因在于高分辨率图像的生成序列像素过大，自注意机制处理受限。为了提升生成图像的分辨率，Zhao 等[86]分两步来生成图像，第一步通过多轴自注意力捕获全局信息来解码空间特征，第二步用多层感知机替代自注意力来减少计算复杂度。此外，文献[87-88]进一步在网络结构上进行改进，力求生成更高分辨率的图像。目前，基于 Transformer 构建的 GAN 成为一大研究热点，但相较于基于 CNN 构建的 GAN 会带来更多计算成本，因此需要寻求更为有效的自注意力形式，从而在性能上进一步提升。

基于 Transformer 的视频图像生成方法依托离散化的处理方式，实现了数据的高效表示。采用自然语言处理的方法实现重建，更好地建立特征的上下文关系，为一些由文本生成视频图像任务建立了有效机制，进而将其应用于跨模态的视频编码与重建。但此类方法计算成本高，难以训练，对于实时视频的应用还有待进一步研究。

6 存在的问题及研究方向

尽管近年来一些生成模型在视频图像重建上取得了显著效果，但现阶段仍存在以下问题亟须解决。

1) 视频长时相关性

视频长时相关性是指不同图像序列之间内容存在较大关联，主要体现在两方面，一是同一视频的长时相关性，其不仅局限于一个 GoP 内的视频帧处理；二是不同时间下视频内容的相似性，如大致相同背景、不同背景下相同人物的视频通话。目前的重建方法集中在消除同一视频的时间相关性上，但对于时间跨度较大的视频帧，仅依据关键帧和边缘、特征点等辅助信息，有时无法保证重建质量。

针对同一视频的长时相关性，可以通过提升辅助信息的提取质量来改善长时视频帧的重建质量，如优化边缘、特征点提取算法，其次针对视频的特定场景来进一步完善重建模型。对于不同时间、不同内容的视频之间存在的重复性内容，则需进一步探究视频图像的语义表征，对内容实现高层概念认知来消除语义冗余，例如，通过对卷积后的高级特征空间进行处理，在编码端和解码端形成语义库，根据特征辨识只传输细节变化的内容，从而大幅度降低传输码率。

2) 高昂的计算成本和时间成本

基于深度生成模型辅助的编码重建框架是以高昂的计算成本为代价来换取编码效率和重建质量。先进的生成模型如 VQ-VAE、GAN 等能够实现清晰的视频图像重建，但这种大型模型的弊端除了消耗巨大的计算和存储资源外，还需要根据大量数据集花费大量时间训练网络模型。

此前，掩码建模方法证明了自监督学习图像表征以及重建的有效性。因此针对此问题，可以将基于掩码视觉标记的方法进一步扩展于视频编码重建中，结合视频的帧间相关性完善掩码学习策略，在一定周期内保证较高的掩码率，并在训练学习时有针对性地跳过掩码区域来减少训练时间和资源。此外，还可以借助小样本学习方法减少数据依赖，借助模型剪枝等压缩方法减小模型参数。

3) 适用场景单一

尽管现有生成模型在视频图像生成任务取得了巨大成功，但将其应用在视频图像编码框架中的研究方法相对较少，且大部分面向图像压缩编码，

对于视频的场景还有待进一步开发。其次，基于特征的重建主要集中在以人物为主的视频场景，无法适用于大部分视频场景，且重建视频的连贯性以及与原视频的一致性还有待提升。此外，目前模型依托大量数据集进行离线训练，并不适用于视频通话等实时业务。

针对此问题，则需寻求更有效的语义表征，设计更先进的重建算法。一方面可以利用图神经网络，基于边和节点的方式刻画物体表征，形成结构化的概念表示，从而实现对复杂物体更灵活通用的建模，同时也更符合人类的视觉感知。另一方面可以加强特征空间的探索，在编码端形成层次特征，根据重建难度选择特征传输等级。对于实时视频业务，可以设计高效的重建算法，利用前几帧视频作为训练样本，结合离线训练模型进行微调，从而完成后续视频帧的重建。

7　结束语

近年来，深度生成模型在视频图像补全、动作迁移、视频图像合成等多个领域取得巨大成功，为视频图像压缩领域的重建模块提供了新的解决方案。本文主要总结了 5 种现有视频图像重建的相关方法，包括传统重建方法及其优化以及 4 种基于生成模型的重建方法，其中重点介绍了生成式的重建方法，根据视频图像数据不同类型的语义表征对模型进行分类、梳理和阐述，最后总结了现有重建方法在视频长时相关性、计算成本和适用场景等方面所存在的问题，探索了相应的解决方案以及进一步的研究方向。

参考文献：

[1] WIEGAND T, SULLIVAN G J, BJONTEGAARD G, et al. Overview of the H.264/AVC video coding standard[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2003, 13(7): 560-576.

[2] SULLIVAN G J, OHM J R, HAN W J, et al. Overview of the high efficiency video coding (HEVC) standard[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 22(12): 1649-1668.

[3] SEGALL A, BARONCINI V, BOYCE J, et al. JVET-H1002: joint call for proposals on video compression with capability beyond HEVC[J]. Joint Video Exploration Team (JVET) of ITU-T SG, 2017, 16: 18-24.

[4] 胡铭菲, 左信, 刘建伟. 深度生成模型综述[J]. 自动化学报, 2022, 48(1): 40-74.
HU M F, ZUO X, LIU J W. Survey on deep generative model[J]. Acta Automatica Sinica, 2022, 48(1): 40-74.

[5] SUN Q, GUO C, YANG Y, et al. Semantic-assisted image compression[J]. arXiv Preprint, arXiv: 2201.12599, 2022.

[6] LUO S H, YANG Y Z, YIN Y L, et al. DeepSIC: deep semantic image compression[C]//International Conference on Neural Information Processing. Berlin: Springer, 2018: 96-106.

[7] KINGMA D P, WELLING M. Auto-encoding variational Bayes[J]. arXiv Preprint, arXiv: 1312.6114, 2013.

[8] CRESWELL A, WHITE T, DUMOULIN V, et al. Generative adversarial networks: an overview[J]. IEEE Signal Processing Magazine, 2018, 35(1): 53-65.

[9] AKAIKE H. Fitting autoregressive models for prediction[J]. Annals of the Institute of Statistical Mathematics, 1969, 21(1): 243-247.

[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. arXiv Preprint, arXiv: 1706.03762, 2017.

[11] 王万良, 李卓蓉. 生成式对抗网络研究进展[J]. 通信学报, 2018, 39(2): 135-148.
WANG W L, LI Z R. Advances in generative adversarial network[J]. Journal on Communications, 2018, 39(2): 135-148.

[12] LI J H, LI B, XU J Z, et al. Fully connected network-based intra prediction for image coding[J]. IEEE Transactions on Image Processing, 2018, 27(7): 3236-3247.

[13] CUI W X, ZHANG T, ZHANG S P, et al. Convolutional neural networks based intra prediction for HEVC[J]. arXiv Preprint, arXiv: 1808.05734, 2018.

[14] HU Y Y, YANG W H, LI M D, et al. Progressive spatial recurrent neural network for intra prediction[J]. IEEE Transactions on Multimedia, 2019, 21(12): 3024-3037.

[15] ZHU L W, KWONG S, ZHANG Y, et al. Generative adversarial network-based intra prediction for video coding[J]. IEEE Transactions on Multimedia, 2020, 22(1): 45-58.

[16] ZHAO L, WANG S Q, ZHANG X F, et al. Enhanced motion-compensated video coding with deep virtual reference frame generation[J]. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society, 2019, 28(10): 4832-4844.

[17] GUO Y, LIU Z Z, CHEN Z Z, et al. Deep inter coding with interpolated reference frame for hierarchical coding structure[C]//Proceedings of 2020 IEEE International Conference on Visual Communications and Image Processing. Piscataway: IEEE Press, 2020: 302-305.

[18] ZHAO Z H, WANG S Q, WANG S S, et al. Enhanced bi-prediction with convolutional neural network for high-efficiency video coding[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(11): 3291-3301.

[19] YAN N, LIU D, LI H Q, et al. Convolutional neural network-based fractional-pixel motion compensation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(3): 840-853.

[20] WANG Z H, CHEN J, HOI S C H. Deep learning for image super-resolution: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3365-3387.

[21] LI Y, LIU D, LI H Q, et al. Convolutional neural network-based block up-sampling for intra frame coding[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(9): 2316-2330.

[22] AFONSO M, ZHANG F, BULL D R. Video compression based on

spatio-temporal resolution adaptation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(1): 275-280.

[23] KIM J, LEE J K, LEE K M. Accurate image super-resolution using very deep convolutional networks[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 1646-1654.

[24] JIANG F, TAO W, LIU S H, et al. An end-to-end compression framework based on convolutional neural networks[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(10): 3007-3018.

[25] SOHN K, YAN X C, LEE H, et al. Learning structured output representation using deep conditional generative models[C]//Proceedings of International Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2015: 3483-3491.

[26] ESSER P, SUTTER E. A variational U-net for conditional appearance and shape generation[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8857-8866.

[27] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 5967-5976.

[28] GREGOR K, DANIHELKA I, GRAVES A, et al. Draw: a recurrent neural network for image generation[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning. New York: PMLR, 2015: 1462-1471.

[29] GREGOR K, BESSE F, REZENDE D J, et al. Towards conceptual compression[J]. arXiv Preprint, arXiv: 1604.08772, 2016.

[30] SØNDERBY C K, RAIKO T, MAALØE L, et al. Ladder variational autoencoders[J]. arXiv Preprint, arXiv: 1602.02282, 2016.

[31] MAALØE L, FRACCARO M, LIÉVIN V, et al. BIVA: a very deep hierarchy of latent variables for generative modeling[J]. arXiv Preprint, arXiv: 1902.02102, 2019.

[32] VAHDAT A, KAUTZ J. NVAE: a deep hierarchical variational autoencoder[J]. arXiv Preprint, arXiv: 2007.03898, 2020.

[33] KINGMA D P, SALIMANS T, JOZEFOWICZ R, et al. Improved variational inference with inverse autoregressive flow[J]. arXiv Preprint, arXiv: 1606.04934, 2016.

[34] OORD A V D, VINYALS O, KAVUKCUOGLU K. Neural discrete representation learning[J]. arXiv Preprint, arXiv: 1711.00937, 2017.

[35] RAZAVI A, OORD A V D, VINYALS O. Generating diverse high-fidelity images with VQ-VAE-2[J]. arXiv Preprint, arXiv: 1906.00446, 2019.

[36] MIRZA M, OSINDERO S. Conditional generative adversarial nets[J]. arXiv Preprint, arXiv: 1411.1784, 2014.

[37] WANG T C, LIU M Y, ZHU J Y, et al. High-resolution image synthesis and semantic manipulation with conditional GANs[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8798-8807.

[38] HU Y Y, YANG S, YANG W H, et al. Towards coding for human and machine vision: a scalable image coding approach[C]//Proceedings of 2020 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2020: 1-6.

[39] KIM S, PARK J S, BAMPIS C G, et al. Adversarial video compression guided by soft edge detection[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2020: 2193-2197.

[40] CHANG J H, MAO Q, ZHAO Z H, et al. Layered conceptual image compression via deep semantic synthesis[C]//Proceedings of 2019 IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2019: 694-698.

[41] 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.

[42] DOLLÁR P, ZITNICK C L. Structured forests for fast edge detection[C]//Proceedings of 2013 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2013: 1841-1848.

[43] XIE S N, TU Z W. Holistically-nested edge detection[C]//Proceedings of 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015: 1395-1403.

[44] HA S, KERSNER M, KIM B, et al. MarioNETte: few-shot face reenactment preserving identity of unseen targets[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 10893-10900.

[45] ZHAO R Q, WU T Y, GUO G D. Sparse to dense motion transfer for face image animation[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Piscataway: IEEE Press, 2021: 1991-2000.

[46] ZAKHAROV E, IVAKHNENKO A, SHYSHEYA A, et al. Fast bi-layer neural synthesis of one-shot realistic head avatars[C]//European Conference on Computer Vision. Berlin: Springer, 2020: 524-540.

[47] FENG D H, HUANG Y, ZHANG Y W, et al. A generative compression framework for low bandwidth video conference[C]//Proceedings of 2021 IEEE International Conference on Multimedia & Expo Workshops. Piscataway: IEEE Press, 2021: 1-6.

[48] NIRKIN Y, KELLER Y, HASSNER T. FSGAN: subject agnostic face swapping and reenactment[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 7183-7192.

[49] OQUAB M, STOCK P, GAFNI O, et al. Low bandwidth video-chat compression using deep generative models[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE Press, 2021: 2388-2397.

[50] PARK T, LIU M Y, WANG T C, et al. Semantic image synthesis with spatially-adaptive normalization[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 2332-2341.

[51] HONG F T, ZHANG L, SHEN L, et al. Depth-aware generative adver-saril network for talking head video generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni-

tion. Piscataway: IEEE Press, 2022: 3397-3406.

[52] SIAROHIN A, LATHUILIÈRE S, TULYAKOV S, et al. Animating arbitrary objects via deep motion transfer[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 2372-2381.

[53] SIAROHIN A, LATHUILIÈRE S, TULYAKOV S, et al. First order motion model for image animation[J]. arXiv Preprint, arXiv: 2003.00196, 2020.

[54] WANG T C, MALLYA A, LIU M Y. One-shot free-view neural talking-head synthesis for video conferencing[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 10034-10044.

[55] KONUKO G, VALENZISE G, LATHUILIÈRE S. Ultra-low bitrate video conferencing using deep image animation[C]//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2021: 4210-4214.

[56] WANG T C, LIU M Y, TAO A, et al. Few-shot video-to-video synthesis[J]. arXiv Preprint, arXiv:1910.12713, 2019.

[57] CHAN C, GINOSAR S, ZHOU T H, et al. Everybody dance now[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 5932-5941.

[58] XIA S F, LIANG K, YANG W H, et al. An emerging coding paradigm vcm: a scalable coding approach beyond feature and signal[C]//Proceedings of 2020 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2020: 1-6.

[59] PRABHAKAR R, CHANDAK S, CHIU C, et al. Reducing latency and bandwidth for video streaming using keypoint extraction and digital puppetry[J]. arXiv Preprint, arXiv: 2011.03800, 2020.

[60] WU Y J, HE T Y, CHEN Z B. Memorize, then recall: a generative framework for low bit-rate surveillance video compression[C]//Proceedings of 2020 IEEE International Symposium on Circuits and Systems. Piscataway: IEEE Press, 2020: 1-5.

[61] SHI X, CHEN Z, WANG H, et al. Convolutional LSTM network: a machine learning approach for precipitation nowcasting[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2015: 802-810.

[62] WANG T C, LIU M Y, ZHU J Y, et al. Video-to-video synthesis[J]. arXiv Preprint, arXiv:1808.06601, 2018.

[63] PAN J T, WANG C Y, JIA X, et al. Video generation from single semantic label map[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 3728-3737.

[64] ZHU P H, ABDAL R, QIN Y P, et al. SEAN: image synthesis with semantic region-adaptive normalization[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 5103-5112.

[65] AKBARI M, LIANG J, HAN J N. DSSLIC: deep semantic segmentation-based layered image compression[C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2019: 2042-2046.

[66] HOANG T M, ZHOU J J, FAN Y B. Image compression with encoder-decoder matched semantic segmentation[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE Press, 2020: 619-623.

[67] CHANG J H, ZHAO Z H, YANG L B, et al. Thousand to one: semantic prior modeling for conceptual coding[C]//Proceedings of 2021 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2021: 1-6.

[68] OORO A V D, KALCHBRENNER N, KAVUKCUOGLU K. Pixel recurrent neural networks[C]//Proceedings of the 33th International Conference on International Conference on Machine Learning. New York: PMLR, 2016: 1747-1756.

[69] OORO A V D, KALCHBRENNER N, VINYALS O, et al. Conditional image generation with PixelCNN decoders[J]. arXiv Preprint, arXiv: 1606.05328, 2016.

[70] SALIMANS T, KARPATHY A, CHEN X, et al. PixelCNN++: improving the PixelCNN with discretized logistic mixture likelihood and other modifications[J]. arXiv Preprint, arXiv: 1701.05517, 2017.

[71] CHEN X, MISHRA N, ROHANINEJAD M, et al. Pixelsnail: an improved autoregressive generative model[C]//Proceedings of the 35th International Conference on International Conference on Machine Learning. New York: PMLR, 2018: 864-872.

[72] MENICK J, KALCHBRENNER N. Generating high fidelity images with subscale pixel networks and multidimensional upscaling[J]. arXiv Preprint, arXiv: 1812.01608, 2018.

[73] KALCHBRENNER N, OORD A, SIMONYAN K, et al. Video pixel networks[C]//Proceedings of the 34th International Conference on International Conference on Machine Learning. New York: PMLR, 2017: 1771-1779.

[74] PARMAR N, VASWANI A, USZKOREIT J, et al. Image transformer[C]//Proceedings of the 35th International Conference on International Conference on Machine Learning. New York: PMLR, 2018: 4055-4064.

[75] ESSER P, ROMBACH R, OMMER B. Taming transformers for high-resolution image synthesis[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 12868-12878.

[76] CHEN M, RADFORD A, CHILD R, et al. Generative pretraining from pixels[C]//Proceedings of the 37th International Conference on International Conference on Machine Learning. New York: PMLR, 2020: 1691-1703.

[77] YAN W, ZHANG Y Z, ABBEEL P, et al. VideoGPT: video generation using VQ-VAE and transformers[J]. arXiv Preprint, arXiv: 2104.10157, 2021.

[78] RAKHIMOV R, VOLKHONSKIY D, ARTEMOV A, et al. Latent video transformer[J]. arXiv Preprint, arXiv: 2006.10704, 2020.

[79] WU C F, LIANG J, JI L, et al. NÜWA: visual synthesis pre-training for neural visual world creation[J]. arXiv Preprint, arXiv: 2111.12417, 2021.

[80] LEE J, KIM D, HAM B. Network quantization with element-wise gradient scaling[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE

Press, 2021: 6444-6453.

[81] BAO H B, DONG L, WEI F. BEiT: BERT pre-training of image transformers[J]. arXiv Preprint, arXiv: 2106.08254, 2021.

[82] HE K, CHEN X, XIE S, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 16000-16009.

[83] XIE Z, ZHANG Z, CAO Y, et al. Simmim: a simple framework for masked image modeling[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 9653-9663.

[84] CHANG H W, ZHANG H, JIANG L, et al. MaskGIT: masked generative image transformer[J]. arXiv Preprint, arXiv: 2202.04200, 2022.

[85] JIANG Y, CHANG S, WANG Z. TransGAN: two pure transformers can make one strong GAN, and that can scale up[J]. Advances in Neural Information Processing Systems, 2021, 34: 14745-14758.

[86] ZHAO L, ZHANG Z Z, CHEN T, et al. Improved transformer for high-resolution GANs[J]. Advances in Neural Information Processing Systems, 2021, 34: 18367-18380.

[87] ZHANG B W, GU S Y, ZHANG B, et al. StyleSwin: transformer-based GAN for high-resolution image generation[J]. arXiv Preprint, arXiv: 2112.10762, 2021.

[88] KARRAS T, LAINE S, AILA T M. A style-based generator architecture for generative adversarial networks[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 4396-4405.

[作者简介]

王延文（1998− ），女，辽宁辽阳人，东北大学博士生，主要研究方向为计算机视觉、视频图像压缩编码。

雷为民（1969− ），男，山西平遥人，博士，东北大学教授，主要研究方向为多媒体智能信号处理、网络多径传输优化和工业实时通信技术。

张伟（1980− ），女，山东济宁人，博士，东北大学讲师、硕士生导师，主要研究方向为多媒体智能信号处理、网络多径传输优化和工业实时通信技术。

孟欢（1998− ），女，辽宁锦州人，东北大学硕士生，主要研究方向为计算机视觉、视频图像压缩编码。

陈新怡（1994− ），女，河北承德人，东北大学博士生，主要研究方向为计算机视觉、视频图像压缩编码。

叶文慧（1991− ），女，山东烟台人，东北大学博士生，主要研究方向为计算机视觉、视频图像压缩编码。

景庆阳（1994− ），女，辽宁沈阳人，东北大学博士生，主要研究方向为计算机视觉、视频图像压缩编码。