ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 41,No. 2,Mar. 2026,pp. 515－542
DOI:10. 16337/j. 1004-9037. 2026. 02. 012

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

深度学习驱动的视频编码:方法、进展与展望

何小海[1], 李鑫磊[1], 魏海涛[1], 毕晓东[2], 聂尧佳[1], 熊志娜[1], 张皓彦[1], 熊淑华[1]

(1. 四川大学电子信息学院,成都 610065;2. 四川大学网络空间安全学院,成都 610065)

摘　要: 随着视频数据量的爆炸式增长,有限的网络带宽和高计算资源需求对视频传输与存储提出了严峻挑战。在此背景下,持续开发高效的视频编码方法以保障在资源受限条件下提供高质量视频服务具有至关重要的理论意义与应用价值。然而,传统混合视频编码框架已逐渐遭遇瓶颈,编码性能的进一步提升越来越困难。近年来,深度学习凭借其强大的非线性拟合与表征能力,为视频编码领域的优化带来了契机。本文对基于深度学习驱动的视频编码技术进行了系统而详细的分析。首先,简要介绍传统编码框架下的视频编码技术,并进一步探讨结合深度学习在帧内/帧间预测等关键模块中的优化;然后,重点讨论了基于深度学习的端到端视频编码框架的发展历程及关键技术路线,并对其性能进行对比分析;最后,进一步介绍深度学习在视频编码领域的重要研究成果,剖析现有技术所面临的挑战和局限性,并对未来视频编码技术的发展趋势进行了展望。

关键词: 视频压缩;深度学习;神经网络;结合学习的视频编码;端到端视频编码

中图分类号: TN919.81　　**文献标志码:** A

引用格式: 何小海,李鑫磊,魏海涛,等.深度学习驱动的视频编码:方法、进展与展望[J].数据采集与处理,2026,41(2):515-542. HE Xiaohai,LI Xinlei,WEI Haitao, et al. Deep learning-driven video coding: Methods, progress, and perspectives[J]. Journal of Data Acquisition and Processing,2026,41(2):515-542.

引　　言

视频作为信息时代的主要传播载体,在信息的获取与传递方面发挥着不可替代的作用,已成为推动电子信息产业发展的重要驱动力。随着互联网技术和终端电子设备的发展,视频技术也广泛应用于医疗、监控、通信和娱乐等领域,如今视频内容主导着在线活动,已占据全球互联网大部分数据流量。然而,视频应用的持续发展以及高分辨率、高动态范围等视频需求的不断提升,促使全球视频数据规模呈现爆炸式增长。这一趋势导致对网络带宽和存储资源的需求急剧增加,对现有网络基础设施带来了前所未有的压力,从而进一步凸显了高效视频处理与压缩技术研究的迫切性。因此,在有限的网络带宽和存储能力下,视频编码技术在视频传输和处理过程中具有至关重要的作用,其通过充分利用视频在空间和时间维度上的冗余特性,在保证用户主观质量的同时对视频数据进行有效压缩,是视频应用可持续发展的重要保障。

在过去的几十年间,为了满足用户对高质量视频日益增长的需求,学术界和工业界共同制定了一系列具有里程碑意义的视频编码标准。其中,以高级视频编码(Advanced video coding,AVC)[1]、高效视频编码(High efficiency video coding,HEVC)[2]和通用视频编码(Versatile video coding,VVC)[3]为代表的采用传统编码框架的视频编码器,通过不断优化编码策略,实现了显著的性能飞跃,每一代新型视

基金项目: 国家自然科学基金(62271336,62211530110);四川省国际科技创新合作/港澳台科技创新合作项目(2024YFHZ0289);TCL科技创新基金(0020506107005);成都市重点研发支撑计划(2024-YF06-00079-HZ)。
收稿日期: 2026-01-09;**修订日期:** 2026-02-26

频编码标准相较于前一代实现约50%的编码性能提升。上述视频编码标准主要采用基于块的混合编码架构,其核心由帧内/帧间预测、变换、量化和熵编码等模块组成,该结构通过对各个模块进行相对独立的优化设计,在长期演进过程中实现了高效的视频压缩,有效推动了视频编码相关领域的进步。然而,传统编码框架基于块的处理方式也逐渐暴露出一定的局限性。首先,由于编码过程中通常以块为处理单元,各块在预测、变换及量化等阶段具有相对独立性,易导致在块边界处产生不连续现象,引发明显块效应,从而影响视频的主观和客观质量[4]。其次,块处理的顺序依赖性也限制了其大规模并行处理的能力,进而制约了视频编码的效率和处理速度。更关键的是,传统编码方法依赖于手工设计的模块,倾向于各个功能模块进行独立优化。这种局部最优策略以及模块间的差异化使得整体编码框架难以进行有效的联合优化,这在一定程度上限制了总体性能的进一步突破。综上所述,尽管传统编码方法极大地提高了视频压缩性能,但其性能的进一步提升逐渐面临瓶颈,且随之而来的编码复杂度的显著增加难以满足日益复杂多样的视频内容以及用户对更高压缩效率、更好视频质量的需求。

近年来,深度学习技术取得了重大进展,在计算机视觉、自然语言处理等领域得到了广泛应用,其展现出的巨大潜力为图像/视频编码领域的发展开辟了新的研究方向。得益于深度学习网络(Deep neural networks,DNNs)强大的非线性表征能力和自适应能力,这些新兴的编码方法能够自动学习图像/视频数据中复杂且具有高层次语义的特征,从而更有效地捕捉数据内在的时空相关性,并且其强大的自适应能力使其能够高效应对多样化的内容和场景,显著提高了模型的泛化能力。基于深度学习的编码方法不仅在提高压缩效率方面取得了突破,同时也更有利于改善人眼主观视觉质量。在深度学习驱动的视频编码方法早期阶段,研究主要聚焦于构建基于深度学习的自编码器网络模型来实现图像数据的压缩,多项研究[5-12]表明,相对于JPEG[13]、JPEG 2000[14]、BPG[15]等传统图像压缩编码方法,基于深度学习的图像编码方法展现出更高的压缩效率。这一突破性进展直接催生了JPEG AI标准的诞生。作为首个基于端到端深度学习架构的图像编码国际标准,JPEG AI的出现代表了一项里程碑式成就[16]。因此,受基于深度学习的图像编码技术的启发,研究者也试图利用深度神经网络来进一步优化视频编码性能[17]。然而,由于视频数据的复杂多样性,将深度学习直接应用于视频编码面临诸多挑战。随着深度学习技术的进一步发展,许多研究工作逐渐将深度学习技术引入传统视频编码框架中,如结合学习的帧内预测、帧间预测和环路滤波算法等[18-24]。这些结合深度学习编码算法的基本原理是在沿用传统视频编码框架的基础上,利用深度学习网络来代替或优化其中的部分模块。虽然这种结合策略能够提升特定模块的效率,但其并未从根本上改变传统编码框架的基础结构,模块间的耦合性不高,因此仍受限于其固有的优化路径,无法实现整体框架性能的根本性突破。因此,为了进一步克服传统编码框架的局限性,提出了完全基于深度学习的端到端编码框架。与传统方法不同,端到端的视频编码框架可以充分利用深度学习的全局优化策略直接针对比特率和质量进行优化,根据视频内容特性自适应地优化编码方式,避免了传统框架中的人工干预和局部最优问题。并且端到端的框架能够在更高的层次上捕捉视频数据的复杂规律和时空相关性,从而实现压缩性能的进一步提高。目前,深度学习驱动的编码方法已成为视频编码领域主流研究方向之一,并且许多现有的基于深度学习的图像/视频编码方法已达到甚至超越传统编码器的压缩性能,有效推动了图像/视频编码领域新标准的制定和发展。现阶段,深度学习驱动的视频编码技术已演进为两大核心范式:结合深度学习的视频编码和基于深度学习的端到端视频编码[25-26]。

本文旨在系统阐述深度学习驱动的视频编码方法及其发展历程。首先,介绍传统的视频编码技术,包括传统视频编码框架以及视频编码标准。其次,讨论结合学习的视频编码技术,分析在传统框架下利用深度学习网络模型对关键模块进行优化的策略。然后,重点分析完全基于深度学习的视频编码框架及发展过程,并对比其与传统编码框架的根本区别。最后,着重探讨视频编码技术未来的研究方

向与面临的核心挑战,并对全文内容进行总结。为了更好地理解视频编码技术路线,可以参考图1所示的视频编码技术脉络结构。

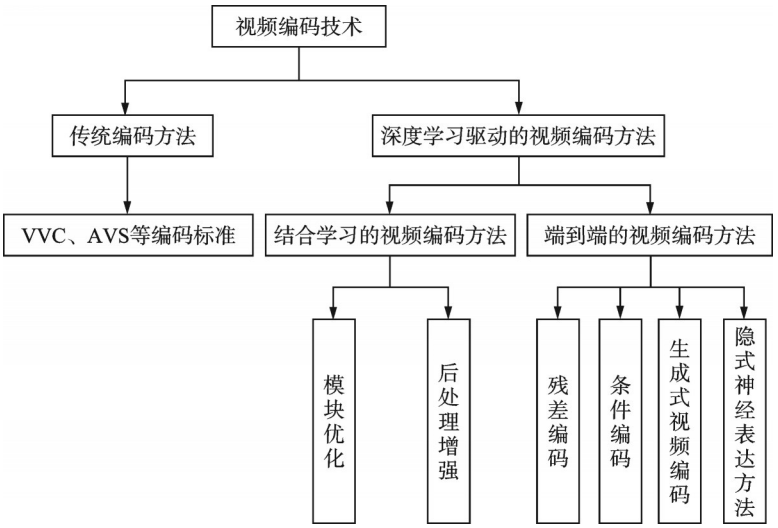

图1 视频编码技术脉络结构图

Fig.1 Structure diagram of video coding technologies

1 传统视频编码技术

数字视频因其数据表示和处理方式,天然存在大量冗余。视频数据中的冗余主要有空间冗余、时间冗余、信息熵冗余、视觉冗余、结构冗余、知识冗余以及统计冗余等。数据冗余是视频可以被压缩的根本原因,视频编码技术的核心目标是通过运用高效的算法来消除这些冗余,从而显著降低传输与存储所需的比特开销,提升数据处理效率与资源利用率[27-28]。

单一的编码工具或技术难以同时处理视频数据中复杂的冗余情况。为此,国际电信联盟-电信标准组织(International telecommunication union-telecommunication standardization sector,ITU-T)在H.261标准中首次提出了混合式编码框架。该框架成功地结合了多种技术来应对不同类型的冗余,不仅实现了压缩效率的显著提升,也为之后所有主流的视频编码标准奠定了技术基础。图2给出了混合式视频编码框架包含的主要模块及步骤。首先,将每个视频帧划分为固定大小的像素块。随后,对像素块进行帧内预测或帧间预测,得到预测值。接着,将原始像素值与预测值相减,得到该像素块的残差。然后,对残差信息进行变换和量化处理。最后,对各个模块的编码模式、参数及残差信息进行熵编码,转换为比特码流,送入信道进行传输。接收端收到码流后,按照与编码过程相反的步骤进行解码,得到重建视频。

在视频编解码发展历程中,不同的组织或研究机构提出了各种标准提案,在这些标准当中,H.26X是国际上主流的视频压缩标准之一。ITU-T于1990年颁布了H.261,1996年颁布了H.263。后来,在H.263的基础上增加了选项,发布了H.263+、H.263++。ITU-T发布的标准主要应用于实时视频通信。为满足实际视频应用对复杂度、实时性和硬件实现的约束,国际标准化组织(International organization for standardization,ISO)制定了MPEG-X系列[29]标准。MPEG-1主要针对存储媒体,MPEG-4更加注重多媒体系统的交互性和灵活性。同时,为了统一视频编码标准,以及满足日益增长的压缩需求,ITU-T和ISO联合起来,先后发布了标准H.262/MPEG-2、H.264/AVC和H.265/HEVC。2015年,两

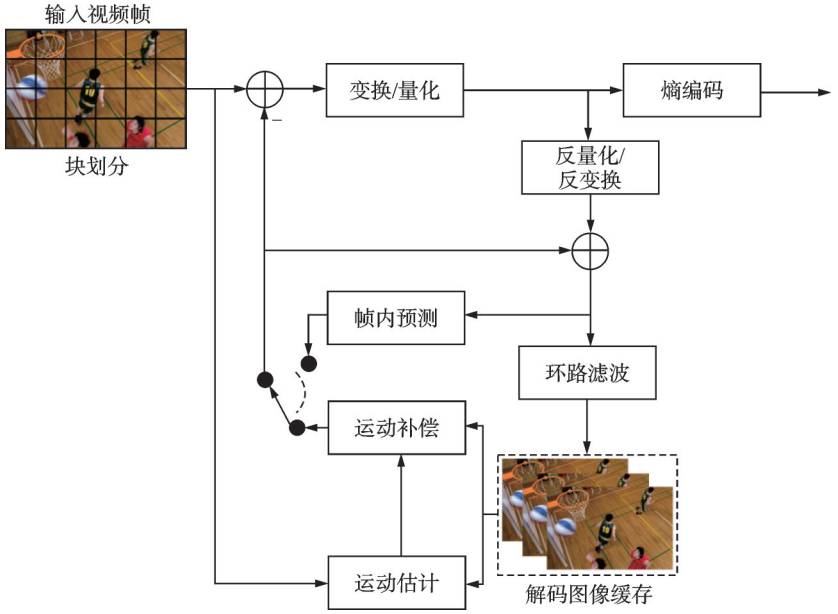

图2 混合式视频编码框架

Fig.2 Framework of hybrid video coding

组织正式成立了联合视频专家组(Joint video experts team,JVET),并在2020年成功完成了最新一代的编码标准H.266/VVC的制定。在相同重建质量的前提下,相较于其前一代视频编码标准,AVC、HEVC和VVC等视频编码标准均达成了约50%的比特率降低。

与此同时,在国内视频编码研究团队中,以高文院士领导的科研团队经过多年的研究,相继推出了AVS(Audio video coding standard)、AVS+/AVS2、AVS3等一系列视频压缩标准。在20年的持续演进中,AVS工作组始终与国际视频编码前沿同步发展,其最新一代标准AVS3在面向8K超高清等新兴场景的压缩性能上已达到甚至部分超越了同期国际主流标准(如H.266/VVC)的水平,尤其是在高分辨率、高动态范围内容的实时编码与自适应码率控制等方面展现出显著优势。图3所示为相关视频编码标准的发展历程。

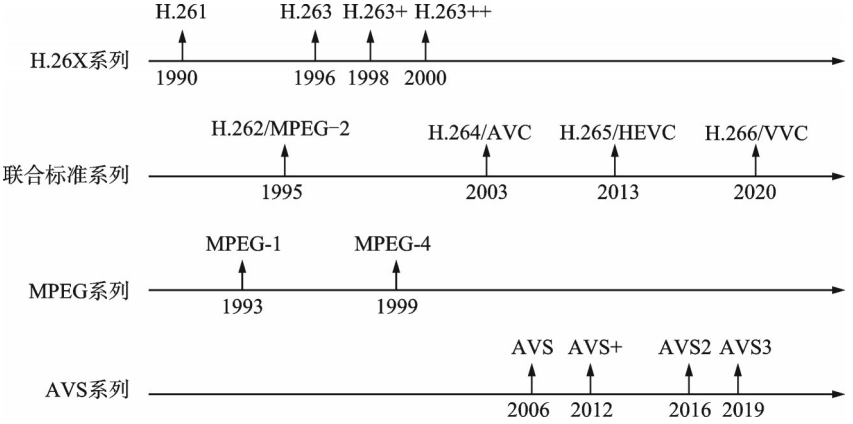

图3 视频编码标准相关发展历程

Fig.3 Evolution of video coding standards

2 结合深度学习的视频编码技术

随着深度学习技术的不断发展,深度学习和传统编码框架相结合的视频编码方法成为新的研究方向。由于传统视频编码器在历经数代标准优化后逐渐面临瓶颈,难以实现性能的进一步提升,所以通过利用深度学习强大的特征提取和学习能力来提高传统编码框架的性能,已成为学术界和工业界共同关注的焦点。总体而言,基于深度学习的混合式视频编码技术主要沿着两条技术路径演进:一是“内部增强”,即基于深度学习的模块优化,将神经网络嵌入编码器内部以改进核心功能;二是“外部修复”,即基于深度学习的后处理增强,在解码端独立运行以提升最终输出质量。这两种路径相辅相成,可共同拓展视频压缩技术的边界。

2.1 结合深度学习的模块优化

基于深度学习的模块优化技术主要是利用神经网络模块来替换或优化传统视频编码中的某些关键模块,从而实现压缩性能的进一步提升。由于传统的混合式编码框架采用模块化设计,每个模块(如预测、变换、量化、滤波)功能相对独立,这为神经网络的针对性集成提供了天然的切入点。通过数据驱动的方式,神经网络能够学习比手工设计模型更复杂的映射关系,从而在各个环节实现更高效的冗余去除。具体地,本节将聚焦于编码框架中的4个核心模块——帧内预测、帧间预测、量化和环路滤波,系统阐述深度学习模型与之结合的最新研究进展。

2.1.1 结合深度学习的帧内预测技术

帧内预测是消除视频帧内空间冗余的核心技术,其基本原理是利用当前编码块周围已重建的相邻像素(通常为上侧和左侧的L形区域),来预测块内待编码的像素值。编码器只需传输预测残差(原始块与预测块的差值)及所选的预测模式,从而大幅压缩数据量。传统方法依赖于一组预设的线性预测模式,如H.265/HEVC中的33种角度模式、DC模式和Planar模式。然而,自然图像内容复杂,固定的角度预测或简单的模板匹配难以准确刻画复杂的纹理结构、精细的边缘方向以及非局部的自相似性,导致预测残差能量仍然较高,限制了压缩效率的进一步提升。

为了突破传统线性预测模型的局限,研究者们转而探索基于深度学习的非线性预测方法。早期开创性工作由Li等[30]完成,他们提出了一种用于帧内预测的深度学习方法。与传统方法使用固定的预测规则不同,他们采用一个完全连接的网络(多层感知机)来学习从相邻重构像素到当前块的端到端映射。该方法的优势在于利用了更广泛的上下文信息(不仅仅是单行/单列像素),并且展现出良好的泛化能力——在特定比特率下训练的模型在其他比特率设置下也能有效工作。与H.265/HEVC相比,该方法可实现平均3.4%的比特率节省,成功验证了使用全连接网络学习空间预测的可行性。然而,全连接网络参数量大,且对输入块尺寸敏感。

后续研究朝着更高效、更精准的架构设计发展。Dumas等[31]设计了一种混合神经网络结构,将卷积神经网络(Convolutional neural network, CNN)与全连接神经网络(Fully connected neural network, FCN)相结合。该架构针对视频编码中不同块尺寸的特点进行了差异化处理:对于大尺寸编码块(如64×64),采用能够有效捕捉空间相关性的卷积神经网络进行预测;而对于小尺寸块(如4×4),则使用参数相对较少的全连接网络以保证效率。这种设计在保持高效计算的同时,提升了对大块内容的预测准确性。实验结果显示,该方法相比基线方案平均取得了0.99%的性能提升。为了解决现有方法在参考像素与当前块空间相关性较弱时(如复杂纹理区域)预测效率不足的问题,Jin等[32]提出了一种基于卷积编码器-解码器网络的帧内预测方法。该方法通过数据驱动的方式,让网络学习参考块特征的内在表示,并依此逐步生成预测块,从而增强了对复杂纹理的建模能力。大量实验验证了其有效性,在Y、Cb和Cr分量上分别实现了约3.41%、3.07%和3.44%的码率节省。

另一类重要思路是模拟人类视觉或图像生成的顺序过程。Hu等[33]设计了一个渐进式空间递归神经网络(Progressive spatial recursive neural network, PS-RNN)来进行帧内预测。该网络由3个空间循环单元组成,通过将已生成部分的信息递归地传递到后续待编码区域,从而逐步生成整个预测块。这种方法更符合图像内容的空间依赖关系。此外,该方法支持可变块大小的帧内预测,在实际编码条件下更加实用。实验表明,在相同的重构质量下,与H.265/HEVC相比,该方案在可变块大小设置下平均降低了2.5%的比特率。Wang等[34]则提出一种基于多尺度卷积神经网络的帧内预测方法,其创新之处在于将传统预测与深度学习优化相结合。首先通过传统角度预测生成一个初始预测块,随后将该初始块与相邻已重建的、范围更广的多行L形参考像素共同输入所设计的网络。该网络利用多尺度特征提取机制,充分融合不同感受野下的上下文信息,同步优化预测块中各个区域的像素值。 在纯帧内编码配置下,相较于H.265/HEVC,所提方法可实现平均3.4%(最高达5.6%)的码率节省,体现了传统方法与深度学习协同增效的潜力。

更具颠覆性的尝试是将生成式模型引入帧内预测,试图实现超越像素填充的"语义级"内容补全。Zhu等[35]提出一种新型帧内预测方法,将帧内预测重新建模为图像修复任务。该方法借助生成对抗网络(Generative adversarial network, GAN),根据已重建的像素信息,对编码块中缺失的部分进行智能推理和补全。学习后的GAN模型被集成至视频编码器与解码器中,并通过率失真优化机制,使其与传统的基于角度的帧内预测模式进行竞争,从而自适应地选择最优预测方式。实验结果表明,所提算法在帧内编码场景中,亮度和色度分量的平均码率节省率分别达到了7.63%和7.65%,展示了生成模型在创造高保真预测内容方面的强大能力。以上几种结合学习的帧内预测方法比较如表1所示。

表1 结合学习的帧内预测技术比较

Table 1 Comparison of learning-based intra prediction techniques

文献	核心思想	关键技术	主要优势	性能表现
30	利用深度学习学习从相邻像素到当前块的端到端非线性映射,突破传统线性预测局限	全连接网络	利用更广泛的上下文信息,展现了良好的泛化能力	平均3.4%的比特率节省
31	针对不同块尺寸进行差异化设计,实现效率与精度的平衡	CNN与FCN的混合结构	对于大块(如64×64),用CNN捕捉空间相关性;对于小块(如4×4),用FCN保证效率	平均0.99%的性能提升
32	增强对复杂纹理和弱空间相关区域的预测能力	卷积编码器-解码器网络	通过编码器学习参考块特征的内在表示,再解码生成预测块	Y、Cb和Cr分量平均节约3.41%、3.07%和3.44%
33	模拟图像生成的顺序过程,更符合空间依赖关系	渐进式空间递归神经网络	通过递归单元逐步生成预测块,同时支持可变块大小预测,实用性高	平均降低2.5%的比特率
34	传统方法与深度学习协同增效	多尺度卷积神经网络	将传统角度预测的初始块和多行L形参考像素作为输入,然后利用多尺度特征提取,融合不同感受野信息,同步优化各区域像素	平均3.4%(最高达5.6%)的码率节省
35	将帧内预测重构为图像修复/内容生成任务	生成对抗网络	进行"语义级"智能推理与补全	亮度/色度分量平均节省达7.63%/7.65%

基于深度学习的帧内预测技术已从早期的可行性验证,发展为包含混合架构、编码器-解码器、递归生成、多尺度融合以及生成式模型在内的多元化技术体系。其核心演进逻辑是让模型从海量数据中自动学习比固定模式更丰富、更灵活的空间先验知识,从而生成更贴近原始图像信号的预测,从根本上降低残差能量。

2.1.2 结合深度学习的帧间预测技术

传统帧间预测技术虽然能有效利用时间冗余,但其基于块匹配和刚性平移运动的假设,在处理复杂运动、形变及光照变化时存在固有局限。这直接导致了预测块不精确、残差能量高,成为制约编码效率进一步提升的关键瓶颈。深度学习技术的引入,为从根本上提升预测精度、突破传统模型限制提供了新的技术路径。研究主要从两个层面展开:一是对传统运动补偿预测结果进行深度增强,二是直接利用深度网络生成高质量的预测信号。

在运动补偿增强方面,Huo等[36]提出的卷积神经网络运动补偿细化方法(CNNMCR)是这一方向的代表性工作。该方法设计了一个CNN网络,以前一帧的运动补偿预测块以及当前帧已重建的上下文区域为输入,直接输出一个优化后的预测块。该网络通过学习,能够有效抑制因量化和运动不准确导致的噪声,并平滑块边界,从而提升预测块的质量。这种对传统预测结果的直接智能后滤波,在低延迟P帧配置下可取得平均2.3%的BD-rate增益。为了更系统地解决复杂运动场景下传统预测残差仍较大的问题,Wang等[37]提出了一个更为精细的三阶段网络框架。该方法不再局限于表面滤波,而是深入估计和补偿预测块与真实内容之间的内在差异,其框架包含:(1)残差估计网络。利用当前块的空间邻域信息初步估计潜在残差。(2)特征融合网络。将初步估计的残差特征与原始预测块的特征进行有效拼接。(3)深度优化网络。对融合后的特征进行深度处理,生成最终精细化的残差图,并将其叠加到原始预测块上得到优化预测。这种分层递进、逐步细化的设计思想,使网络能够自适应地处理不同复杂程度的运动不一致性问题,因此在低延迟P帧、低延迟B帧及随机接入等多种编码配置下均表现优越,分别实现了4.6%、3.0%和2.7%的BD-rate增益。

在直接生成预测信号方面,研究者们致力于超越传统运动补偿范式。Yan等[38]将分数像素运动补偿构建为一个回归问题。他们摒弃了传统固定系数的插值滤波器,转而采用一个可学习的卷积神经网络。该网络以参考帧的整数像素点及分数运动向量为输入,直接回归出高精度的亚像素值。这种方法使网络能够从数据中自适应地学习最优的插值核函数,从而生成更精确的亚像素参考值,显著提升了运动补偿的精细度。实验证明,这种数据驱动的分数像素生成方法在多种编码配置下均能带来显著的码率节省。

更具创造性的尝试是直接生成全新的参考内容。这类方法旨在为编码器提供传统方法无法获得的、质量更高的预测选项。Zhao等[20]提出的基于CNN的帧率上转换预测方法即属此类。该方法的核心是解决双向预测中虚拟中间帧生成质量不高的问题。它利用深度学习强大的时空建模与内容生成能力,从前后两个已重建的参考帧中,合成出一个符合真实运动轨迹的高质量中间帧块作为预测。这实质上为编码器增加了一个基于深度学习的“虚拟双向预测”模式,通过率失真优化与传统模式竞争,取得了平均超过3%的编码增益。Lin等[39]则将这一思路推向极致,致力于解决长期运动建模和外推帧真实性的挑战。他们利用生成对抗网络直接进行前向帧外推。通过精心设计的拉普拉斯生成对抗网络金字塔结构,该模型能够基于过去若干帧的历史信息,直接推理并生成未来的一帧完整图像,并将其作为额外的参考帧引入编码环路。这不再是简单的运动轨迹插值,而是为编码器增加了一个具有强语义理解和内容生成能力的“智能参考帧”。这种基于深度生成模型的帧外推方法,在低延迟P帧配置下平均可实现2.0%的BD-rate压缩增益。以上几种结合学习的帧间预测方法比较如表2所示。

这些研究共同表明,深度学习不仅能够优化传统帧间预测流程的各个环节,更能创造新的预测机

表2 结合学习的帧间预测技术比较

Table 2 Comparison of learning-based inter prediction techniques

文献	核心思想	关键技术	主要优势	性能表现
36	对传统运动补偿预测块进行智能后滤波,抑制噪声、平滑边界,直接提升预测块质量	CNN网络	输入是前一帧的MC预测块和当前帧已重建的上下文区域。输出是优化后的预测块	低延迟P帧下,平均1.8%(Y)和2.3%(UV)的BD-rate增益
37	系统性地估计并补偿预测块与真实内容的内在差异,处理复杂运动导致的大残差问题	三阶段网络	残差估计网络:初步估计潜在残差;特征融合网络:融合残差与原始预测特征;深度优化网络:生成精细化残差图	低延迟P/B帧及随机接入配置下,分别实现4.6%、3.0%和2.7%的BD-rate增益
38	将分数像素运动补偿构建为数据驱动的回归问题,学习最优插值核,直接生成高精度亚像素值	可学习的CNN	超越传统固定系数插值滤波器	在多种编码配置下均带来显著的码率节省
20	利用深度学习合成高质量的虚拟中间帧块,为双向预测提供新的、更优的预测选项	CNN模型	从前后两个参考帧合成高质量中间帧块,作为新增的"虚拟双向预测"模式参与RDO竞争	平均取得超过3%的编码增益
39	为编码器生成具备强语义理解能力的"智能参考帧",实现超越传统运动轨迹插值的长期内容生成	拉普拉斯GAN金字塔结构	基于历史帧信息,直接推理并生成未来的一帧完整图像,引入编码环路作为额外参考帧	低延迟P帧配置下,平均实现2.0%的BD-rate压缩增益

制。从增强已有预测、改进插值方法,到生成全新参考内容,深度学习正推动帧间预测从基于简单运动模型的匹配,向基于语义理解和内容生成的智能预测演进。这种演进不仅直接降低了时间域的残差能量,更预示着视频编码技术向更高效、更智能方向发展的趋势。

2.1.3 结合深度学习的量化技术

量化模块紧随变换之后,是实现比特压缩的核心有损步骤,其本质是对变换后得到的频域系数进行"取整"操作,将每个系数映射到有限的量化级别上。这个过程会主动舍弃人眼不敏感的高频细节和微弱的能量值,从而在可接受的失真范围内实现数据量的锐减。传统量化通常采用固定的量化步长,由量化参数(Quantization parameter, QP)控制,对所有内容区域进行"一刀切"的处理,未考虑人眼视觉特性的空间差异性。将深度学习方法融入量化过程,旨在引入"感知重要性"或"视觉显著性"的指导,实现内容自适应的智能比特分配。这种非均匀的、感知驱动的量化策略,能够更优地分配有限的比特资源。

早期研究侧重于将视觉显著性检测模型与传统编码决策进行结合。Xiong等[40]提出了一种面向HEVC的显著性感知快速帧内编码算法。他们采用自底向上的视觉显著性计算模型,将显著图离散化后用于指导编码单元(Code unit, CU)大小的快速选择,并对不同显著度的CU分配不同的量化参数偏移量(Delta QP)。实验表明,该算法在平均码率上降低了2.18%,但峰值信噪比(Peak signal to noise ratio, PSNR)有轻微下降(0.15 dB)。Jiang等[41]进一步利用马尔可夫链和光流进行时空显著性检测,并将显著值归一化后用于动态调整每个CU的QP,实现了平均4.26%的码率降低,同时保持了相近的主观质量。更进一步的研究尝试结合恰可察觉失真(Just noticeable distortion, JND)模型,探索人眼感知的失真上限以实现极限压缩。Wu等[42]提出了一种基于JND的QP优化模型,该模型利用集成学习和

轻量级MobileNetV2网络,结合注意力机制和ConvLSTM,从关键帧中预测在JND阈值内的QP值。尽管作者报告了显著的码率降低(超过50%),但超过一半的观看者认为该方法导致了非常明显的视频质量下降,这揭示了JND模型在实际复杂场景下应用的挑战。

另一条更系统的技术路线是将显著性深度整合到率失真优化(Rate distortion optimization, RDO)这一编码核心决策过程中。Zhu等[43]提出了一个基于显著性的HEVC视频压缩优化框架。他们设计了一个先进的视频显著性检测器,结合预训练Res2Net-50提取的多尺度静态特征和时序循环模块提取的运动信息。生成的显著图被用于修改HEVC的拉格朗日率失真优化模型,通过调整失真项的权重来鼓励在视觉重要区域进行更精细的编码。此外,他们还提出了两种显著性引导的QP调整方法。Li等[44]则在VVC框架下提出了一种显著性引导的CU分割与量化控制方案。他们采用一种无需光流计算的时空显著性检测模型,并基于检测结果自适应调整四叉树加多类型树(Quad-tree plus multi-type tree, QTMT)分割的深度,并决定是否跳过某些编码模式(如帧内子分区),同时在CU层级应用分层量化控制。该方法实现了平均3.68%的码率降低。以上几种结合学习的量化方法比较如表3所示。

表3 结合学习的量化技术比较

Table 3 Comparison of learning-based quantization techniques

文献	核心思想	关键技术	主要优势	性能表现
40	早期探索,将视觉显著性与编码决策结合,实现感知驱动的比特分配	自底向上的视觉显著性计算模型	指导CU大小快速选择并根据显著度为不同CU分配不同QP偏移量	平均码率降低2.18%,但PSNR下降0.15 dB
41	通过更精细的时空显著性检测,动态调整每个CU的量化参数	马尔可夫链和光流	利用显著值动态调整每个CU的QP,实现内容自适应的量化	平均码率降低4.26%,同时保持了相近的主观质量
42	探索人眼感知的失真上限(JND),以实现极限压缩	集成学习+轻量级MobileNetV2网络	旨在生成在恰可察觉失真阈值内的QP,追求极致压缩	报告码率降低超50%,但超过半数观看者认为质量下降非常明显,揭示了JND模型应用的挑战
43	将显著性深度整合到率失真优化核心决策中,系统性地优化压缩	结合Res2Net-50(静态)与时序循环模块(动态)	在RDO层面鼓励重要区域精细编码,提供了系统性的感知编码优化框架	实现了基于显著性的优化,提升了视觉质量并降低了编码复杂度
44	在最新的VVC标准下,实现从CU分割到量化控制的全流程显著性引导	无需光流的时空显著性检测模型	自适应调整QTMT分割深度,决定是否跳过某些编码模式,在CU层级应用分层量化控制	平均码率降低3.68%,展示了显著性在VVC复杂编码工具中的引导潜力

总而言之,基于深度学习的量化技术正从简单的、基于显著图的QP调整,向与RDO深度结合、并协同指导CU分割、模式选择等高级编码决策的方向演进,其目标是建立一套以人眼感知质量为最终目标的、内容自适应的智能码率分配体系,推动压缩性能向感知最优的方向发展。

2.1.4 结合深度学习的环路滤波技术

环路滤波是作用于解码重建帧上的后处理环节,是视频编码框架中实现“滤波-预测”良性循环的关键,其主要任务是修复由块划分、量化和运动补偿等过程引入的视觉失真,如块效应、振铃效应和模糊。滤波后的帧不仅用于当前帧的显示,更作为后续帧进行帧间预测的参考帧。一个更干净、更准确的参

考帧能显著提升后续预测的精度,从而在整体上提升编码效率。传统方法(如去块滤波、样本自适应偏移SAO)依赖于固定的或基于局部统计的线性滤波器,其修复能力有限。将深度学习方法引入环路滤波,标志着该模块从依赖手工设计模型到数据驱动智能优化的范式转变。

Huang等[45]提出了一种基于量化参数(QP)可变的卷积神经网络环路滤波器,用于VVC帧内编码。为了避免为不同QP值训练和存储多个独立网络带来的开销,他们创新地设计了QP注意力模块(QP attention module, QPAM)。该模块能够感知输入重建帧所对应的QP值,并在通道维度上对不同QP对应的特征进行自适应加权,从而使单个网络具备处理不同压缩强度失真的能力。实验结果表明,该方法在全帧内配置下平均可实现4.03%的BD-Rate节省,性能甚至优于需要训练多个独立QP网络的方案。

随着Transformer在视觉任务中的成功,研究者开始探索其捕获长程依赖的优势用于环路滤波。Zhang等[46]提出一种融合残差块与Transformer的神经网络环路滤波器,命名为RTNN。该网络通过结合CNN残差块(擅长提取局部特征)与Transformer模块(擅长捕捉非局部关联),能够有效处理VVC中复杂的压缩失真。此外,RTNN中设计了新型注意力模块,并采用多阶段训练策略以充分考虑量化参数的影响。将RTNN集成至VVC测试模型后,实验结果表明,在全帧内和随机接入配置下,该滤波器在Y、Cb、Cr分量上均取得了显著的BD-rate压缩增益。

为了更精细地处理不同特征的失真,更复杂的网络架构被提出。Liu等[47]提出一种基于双查询Transformer的内容自适应神经网络环路滤波器,命名为DQT-CALF。该方法的核心是采用基于双查询机制的Transformer,分别处理代表低频全局信息的查询和代表高频局部信息的查询,从而学习更丰富的特征表示。DQT-CALF包含特征提取、特征增强与重建3个部分,在特征增强阶段设计了多类型特征融合模块,从频域和空间角度对特征进行划分、处理和相互转换。实验结果表明,DQT-CALF相较于基准模型,在VVC的全帧内与随机接入配置下均实现了卓越的BD-rate增益(Y分量8%~9%,U/V分量约22%)。以上几种结合学习的环路滤波方法比较如表4所示。

深度环路滤波技术已成为深度学习在视频编码中应用最成功、最成熟的领域之一。这些先进的网

表4 结合学习的环路滤波技术比较

Table 4 Comparison of learning-based in-loop filtering techniques

文献	核心思想	关键技术	主要优势	性能表现
45	设计单网络多QP处理方案,避免为不同压缩强度训练多个独立模型	CNN+QP注意力模块	QPAM模块:感知输入帧的QP值,在通道维度上对不同QP的特征进行自适应加权	全帧内配置下,平均实现4.03%的BD-rate节省,效率优于多网络方案
46	融合CNN的局部特征提取与Transformer的非局部关联建模能力,处理复杂压缩失真	残差块与Transformer融合网络	首先,CNN残差块提取局部特征,然后使用Transformer模块捕捉长程依赖,形成新型注意力模块与多阶段训练策略	在全帧内和随机接入配置下,Y、Cb、Cr分量上均取得显著的BD-rate增益
47	通过双查询机制实现内容自适应的频带分离与精细处理	基于双查询Transformer的架构	使用双查询机制分别处理低频全局查询与高频局部查询,然后使用多类型特征融合模块,从频域和空间角度划分、处理并转换特征	在VVC全帧内与随机接入配置下,Y分量节省8%~9%,U/V分量节省22%的BD-rate

络能够智能地理解内容并修复失真,其带来的显著编码性能提升,使其成为新一代编码标准中极具竞争力的增强工具。未来的方向包括设计更低复杂度的实时滤波网络,以探索滤波与编码其他环节(如预测)的联合优化。

2.2 基于深度学习的后处理技术

后处理模块是视频编码框架中位于解码端之后的独立增强组件,其核心使命是在码流完全解码后,对最终输出的重建视频序列进行视觉质量提升,其过程不影响编码端的压缩决策与环路内的参考帧内容。与集成在编码器内的环路滤波不同,后处理是完全独立的,不参与编码循环,因此其算法设计与部署具有更大的灵活性。该模块专注于修复并超越传统解码器的重建能力,针对压缩过程中不可避免产生的全局性失真,如细节丢失、整体性模糊、色彩失真以及残留的块效应与噪声等,进行综合性的智能修复与增强。

将深度学习方法应用于后处理,能够解锁前所未有的画质优化潜能。基于海量数据训练的深度神经网络,能够深刻理解视频内容的语义结构与自然图像先验,从而执行从低质量重建帧到高质量视觉呈现的复杂非线性映射。它不仅能更彻底地消除各类编码伪影,更能智能化地“想象”并复原出符合人眼感知的高频纹理、锐利边缘与生动色彩,甚至在超分辨率、色彩增强等维度上超越原始分辨率与色域的限制。这一独立的增强环节为用户提供了在任意已有码流上直接实现主观视觉体验跃升的可能,是提升端侧播放质量的关键技术路径。当前的研究主要致力于设计更强大的网络结构以利用时空信息。

为了充分挖掘视频帧间与帧内的上下文信息,Zhang等[48]提出一种新型多尺度互通时空网络。该网络基于多尺度特征互通模块和源特征选择增强模块构建。多尺度特征互通模块允许不同尺度的特征进行交互,并利用通道统计信息生成注意力权重以重新校准特征;源特征选择增强模块则利用目标帧的浅层特征来指导深层特征的融合,从而实现更精确的像素级预测。生成对抗网络因其能生成高度逼真细节的特性,被广泛用于追求更高感知质量的后处理任务。Wang等[49]提出一种基于多级小波包变换的生成对抗网络(多级小波生成对抗网络)。该网络首先通过金字塔式运动补偿获取时序信息,然后设计一个包含小波密集残差块的重建网络来恢复高频细节,并通过在小波包变换域引入对抗损失,进一步促进视频帧中高频细节的逼真恢复。

一个关键且富有潜力的研究方向是充分利用编码过程中产生的“边信息”或“编码先验”,如运动矢量、模式信息、量化参数等。这些信息为后处理网络提供了至关重要的时空线索。于海等[50]深入研究了视频编码机制与压缩视频质量增强任务之间的内在关联,并提出一种基于三维卷积神经网络(3D-CNN)的非对齐压缩视频质量增强算法。该方法摒弃了对严格帧间对齐的依赖,直接在原始压缩域中建模时空冗余,充分挖掘帧内与帧间的上下文信息。通过3D卷积结构,网络能够联合捕捉空间细节与时间动态,从而实现对压缩失真的精准补偿。在H.265/HEVC低延迟(Low delay, LD)配置下、量化参数为37时,该算法相较标准解码器提升峰值信噪比达0.465 2 dB。Zhu等[51]提出了编码先验引导聚合网络,旨在充分利用这些先验中的时空信息。该网络主要由帧间时序聚合模块与多尺度非局部聚合模块构成,分别用于融合连续帧的时序信息和捕获全局空间信息。尤为重要的是,为促进该领域研究,他们构建了一个包含300段视频的新数据集,其中每段视频都从对应码流中提取了多维编码信息,有效解决了现有数据集缺乏编码先验的缺陷。以上几种基于学习的后处理方法比较如表5所示。

基于深度学习的后处理技术为提升现有视频服务的终端观看体验提供了一条实用且有效的路径。它能在不改变已生成码流的前提下,在播放端对视频进行智能增强。随着网络结构的不断演进和对编码先验信息的深入利用,后处理技术有望为用户带来接近甚至超越原始质量的观看体验。

尽管结合学习的视频编码技术通过引入深度学习技术显著提升了编码性能,但其技术路线本身存在若干局限性。首先,将深度学习模块嵌入传统编码架构,各模块(如帧内预测、帧间预测、环路滤波)通

表5 基于学习的后处理技术比较

Table 5 Comparison of learning-based post-processing techniques

文献	核心思想	关键技术	主要优势	性能表现
48	设计多尺度特征交互与源特征引导机制，以充分利用上下文信息进行精确的像素级预测	多尺度特征互通模块＋源特征选择增强模块	实现了跨尺度的特征交互与自适应校准，并通过浅层特征引导，增强了预测的准确性	实现了更精确的像素级预测，充分挖掘了帧间与帧内的上下文信息
49	利用小波包变换域的对抗学习，专门促进高频细节的逼真恢复，以追求更高的感知质量	金字塔式运动补偿＋波密集残差块重建网络＋小波包变换域	在频域（小波包变换域）引入对抗损失，能更有效地引导网络恢复人眼敏感的高频纹理和细节	有效恢复了视频帧中的高频细节，显著提升了视频的主观感知质量
50	探索视频编码机制与压缩视频质量增强之间的内在关联，提出一种无需帧对齐、直接利用压缩域信息的质量增强方法	基于三维卷积神经网络（3D-CNN）的非对齐重建架构	摒弃传统帧间对齐步骤，在原始压缩域中联合建模时空冗余，有效挖掘帧内与帧间的上下文信息，实现对压缩失真的精准补偿	在H.265/HEVC低延迟（LD）配置、QP＝37时，相比标准解码器PSNR提升0.465 2 dB
51	充分利用编码过程产生的“边信息”（如运动矢量、模式、QP等），为后处理网络提供强时空线索	帧间时序聚合模块＋多尺度非局部聚合模块	开创性地系统化利用编码先验指导后处理，同时构建包含丰富编码信息的数据集，解决了该研究领域的数据瓶颈	通过聚合编码先验中的时空信息，有效提升了后处理效果，并促进了该方向的研究

常独立设计和优化，缺少模块间的全局联合优化，模块间的协同增益往往未能充分发挥。其次，该框架本质上依附于传统编码标准的架构体系，其设计范式受限于现有框架的约束，难以突破传统架构的固有瓶颈，在一定程度上限制了编码范式的根本性创新。此外，深度学习模块与传统编解码模块之间存在接口不匹配、信息传递损失等问题，理论上的性能增益在实际系统应用中往往难以完全兑现。上述局限性表明，混合框架虽在近中期具备较强的实用价值，但其长期演进空间可能受制于对传统架构的路径依赖。

3 基于深度学习的端到端视频编码技术

近年来，随着深度学习理论的不断深入与蓬勃发展，神经视频编解码（Neural video codec，NVC）已逐渐成为突破传统编码性能瓶颈的关键技术。与传统混合编码框架不同，NVC通过非线性表达能力与联合优化机制，能够更好地实现对视频特征的高效提取与时空冗余的自适应消除。根据上下文建模与优化策略的不同，现有的端到端视频编码主流研究可归纳为以下几种框架：基于残差编码的视频压缩、基于条件编码的视频压缩、基于生成式模型的视频压缩以及基于隐式神经表达的视频压缩。这些方法进一步推动了编码框架的复杂度优化，为实现更高效的视频压缩提供了新的方向。

3.1 基于残差编码的端到端视频编码

基于残差编码的视频压缩框架是NVC发展的早期形式，其设计思想很大程度上借鉴了传统混合编码架构，即通过预测当前帧与参考帧之间的残差值来消除时域冗余。2019年，Lu等[52]提出了第一个完全基于深度学习的端到端视频编码压缩框架（Deep video compression，DVC）。DVC首次将传统视频编码的所有关键模块全部替换为卷积神经网络组件，实现了完全的端到端训练。如图4所示，DVC用光流网络替代了传统的块匹配运动估计，以当前帧x_t与前一帧重建帧$\hat{x}_{t-1}$为输入，估计出两帧之间

的运动信息v_t,为了减少传输运动信息的比特数,使用一个运动向量编码网络将运动信息v_t压缩为潜在表示m_t,运动向量解码网络从量化后的$\hat{m}_t$中重建出运动信息$\hat{v}_t$,随后利用运动补偿网络根据前一帧重建帧$\hat{x}_{t-1}$和重建的运动信息$\hat{v}_t$生成预测帧$\hat{x}_t$,预测帧与当前帧在像素域相减得到残差r_t,残差编码网络将r_t映射为非线性的潜在表示$\hat{y}_t$,后又通过残差解码网络得到重建残差$\hat{r}_t$。为了实现端到端的率失真优化,框架引入了比特率估计网络,整个网络通过一个单一的损失函数进行端到端训练,该损失函数综合考虑了失真和比特率,以实现最优的压缩效率。实验表明,DVC在PSNR上优于H.264/AVC,在多尺度结构相似性(Multi scale structural similarity index measure,MS-SSIM)指标上甚至超越了H.265/HEVC。然而,DVC作为基于单参考帧预测的视频编码模型,存在明显的局限性,其无法利用多帧时序信息,且简单的减法运算难以处理复杂的纹理变化和遮挡区域,这使得DVC在一些特殊场景中无法充分利用已解码重建帧的信息,从而造成编码性能的下降。为了解决这些问题,后续研究在DVC基础上进行了大量改进。王惠之等[53]针对卫星遥感场景下预测帧与参考帧未对齐、比特分配不佳等问题,提出内容自适应驱动的深度视频压缩算法,通过多尺度残差压缩、运动矢量预测及内容自适应掩码模块优化比特分配与重建质量,将残差压缩转移至尺度空间,在UVG、HEVC等数据集上表现优于H.264/AVC、H.265/HEVC及主流深度压缩方法,且模型更轻量化。Lin等[54]则在DVC的基础上提出了基于多帧预测的全网络视频压缩模型M-LVC,该模型利用之前的多个相邻帧及其运动信息来预测当前帧及其运动信息,证明了利用历史帧集合作为参考列表能有效抑制预测误差累积,使其在PSNR指标上全面领先于H.265/HEVC。Agustsson等[55]则针对光流网络做出了优化,提出将光流网络推广到尺度空间,通过在不同尺度上建模运动模糊,解决了传统光流难以处理模糊和大幅运动的问题,其指标MS-SSIM在UVG和MCL-JCV数据集上取得了超越H.265/HEVC的性能,且比同期SOTA方法DVC节省了大量码率。

尽管DVC等都是完全基于深度学习的视频编码模型,但都只是简单地用神经网络模块去替换传

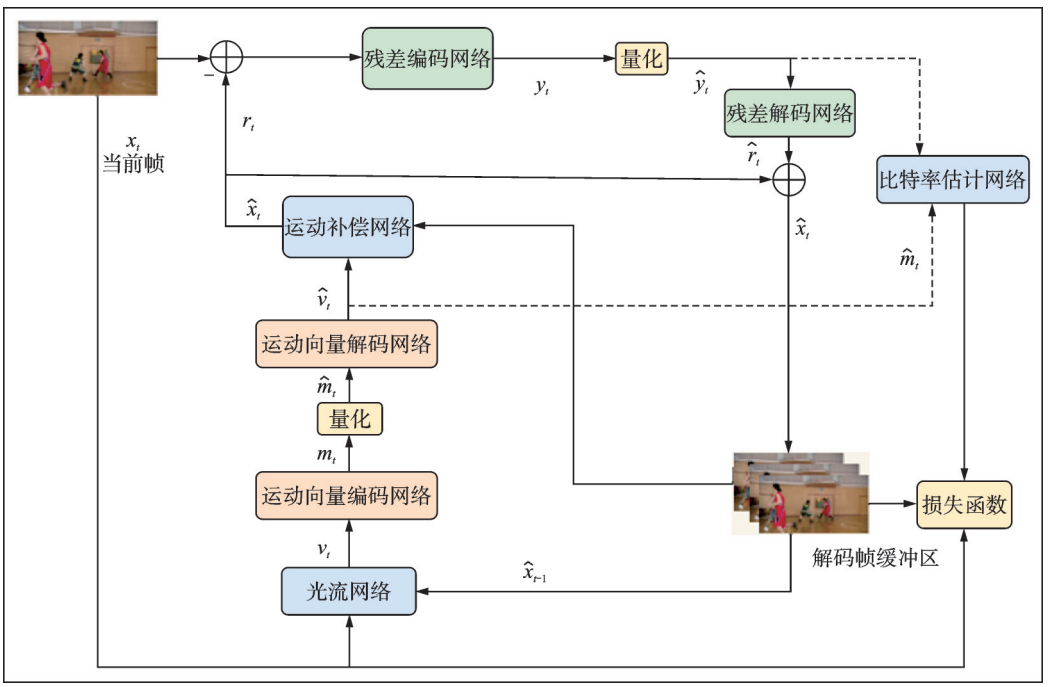

图4 DVC视频编码框架[52]

Fig.4 Framework of DVC video coding[52]

统视频编码框架中对应的功能模块,并未对原有框架有所突破。为了进一步提升基于深度学习的视频编码模型的编码性能,一些研究人员开始突破原有的视频编码框架。传统的像素域残差编码忽略了视觉信号的高维语义特性,因此,Hu等[56]提出了基于特征空间的视频压缩模型FVC,FVC不再在像素空间计算残差,而是先通过特征提取网络将视频帧映射到高维特征空间,在特征空间中进行运动补偿、残差计算和压缩。由于特征空间具有更强大的非线性表达能力,FVC能够容忍不完美的运动估计,并通过可变形卷积在特征层面对齐细节,极大地提升了编码性能。徐智慧等[57]为解决现有编码未充分利用视频帧像素空间与上下文依赖性的问题,提出了基于可变形帧和自回归通道预测的视频压缩算法,通过可变形卷积生成预测帧、通道分割自回归预测优化重建,提升压缩效率与图像质量。而针对可变形卷积缺乏显式运动引导而导致的偏移量溢出及预测帧质量下降问题,张秋建等[58]提出了一种结合卷积神经网络的运动特征提取模块,以挖掘帧间时域冗余并提升运动表示能力。由于运动估计往往导致图像结构的扭曲,因此Gao等[59]提出了一种结构保持的运动估计方法,通过非对称结构的同时参考前向和后向帧来构建运动场,更好地保持了视频的结构信息,这使得FVC等基础架构在标准测试集上实现了约10%的BD-rate性能提升。另外,Liu等[60]提出了多模式视频压缩,其认为视频中的运动模式是多样的,单一的预测方式无法覆盖所有情况,因此在特征域中设计了多种预测模式,并自适应地选择最优模式来描述运动向量,从而更精确地捕捉复杂运动,其在UVG、MCL-JCV等数据集上的率失真表现已展现出抗衡HVVC的卓越竞争力。为了突破早期光流网络的局限,Wang等[61]提出了异构可变形卷积来替代标准的光流网络,通过自适应地学习采样点的偏移量,实现了更灵活的运动补偿,有效解决了传统方法在物体边缘处的对齐难题,其在HEVC数据集上较H.265/HEVC实现了22.08%的码率节省。Zhang等[62]通过设计的时序特征提取器和空间特征提取器分别提取相邻参考帧的时序信息与目标帧的空间信息,然后利用自适应加权特征融合模块自适应地融合这两种信息,并通过多通道增强残差块进一步优化特征以增强重建能力,在QP=37的严苛条件下相比H.265/HEVC基准实现了平均1.02 dB的PSNR增益。

在熵编码方面,为了更准确地估计潜在变量的概率分布,Cheng等[63]提出使用高斯混合似然来参数化潜在变量的分布,相比单一高斯模型,能更灵活地拟合复杂的数据分布,从而减少熵编码的比特消耗。针对端到端模型的码率控制,Rippel等[64]提出灵活的码率控制方法,其可以根据不同场景进行码率调节,并设计了一个高效的主干网络以减少编码的复杂度,在NVIDIA Titan V上实现了1 080P视频18 fps的实时解码,并在UVG数据集上相比H.264/AVC节省了44%的BD-rate。

3.2 基于条件编码的视频编码

近年来,基于残差编码框架的视频压缩取得了显著的进展,然而,根据信息论,条件熵总是小于或等于残差熵,因此条件编码在理论上具有更高的压缩上限。如图5所示,与残差编码框架相比,条件编码框架不再局限于单纯地处理当前帧的像素级残差,而是将经过运动补偿的相邻参考帧特征作为先验条件,与当前帧的特征分布进行联合建模[65]。通过这种机制,模型能够在特征空间内构建全局性的时空上下文,从而学习到当前帧相对于参考帧的条件概率分布,这种基于条件概率的建模方式显著增强了模型对视频内容的预测精度。这使得条件编码框架在面对不同类型的视频数据时,能够根据具体场景进行优化,从而实现更加高效的压缩和更好的视频质量。此外,条件编码框架在系统设计上具备高度的可扩展性与灵活性,其允许根据视频内容的语义复杂度和目标比特率约束,动态调整特征提取与熵模型的编码策略,这种内容自适应的特性使得模型能够针对不同类型的视频数据进行细粒度的优化,最终在保证视觉重建质量的同时,实现率失真性能的最大化。

Li等[65]开创性地提出了深度上下文视频压缩(Deep contextual video compression,DCVC)框架。该框架标志着神经视频编码从预测编码到条件编码的转变,DCVC将特征域的上下文信息作为条

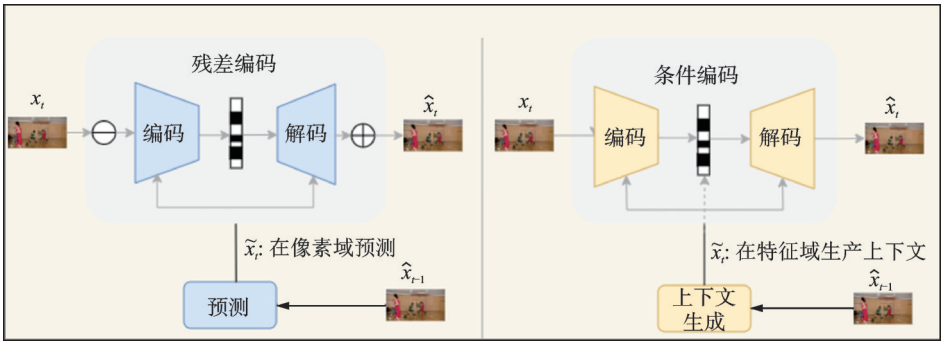

图5 基于残差编码与条件编码的视频压缩框架对比[65]

Fig.5 Comparison of video compression frameworks based on residual and conditional codings[65]

件先验,通过学习高维特征的条件概率分布来指导当前帧的重建。实验验证,DCVC在1 080P标准测试集中相比x265实现了26.0%的码率节省。针对DCVC中上下文信息利用不足的问题,Sheng等[66]提出了DCVC-TCM,该模型引入了多尺度时域上下文挖掘机制,提取并在多个模块间传播长距离的时序依赖特性。为了提升推理速度,DCVC-TCM摒弃了计算昂贵的自回归空间先验,转而利用特征传播来维持性能,相比H.265/HEVC,该方案实现了14.4%的码率节省;在MS-SSIM评价体系下,其相比H.266/VVC更是实现了21.1%的性能领先,证明了深度挖掘特征级上下文对提升预测精度的关键作用。由于去除自回归先验可能会导致熵编码性能下降,Li等[67]进一步提出了DCVC-HEM,该工作设计了一种混合时空熵模型,能够同时捕捉空间和时间维度的统计依赖性,并引入多粒度量化机制。实验表明,该模型在UVG数据集上相比H.266/VVC最高配置实现了18.2%的BD-rate节省。为了进一步挖掘冗余信息,Li等[68]提出了DCVC-DC,该框架通过引入多样化上下文策略,利用多个参考帧及自适应权重优化当前帧的特征表示,此外,由于采用四叉树的高效熵编码方法,DCVC-DC成功超越了传统编码标准ECM的性能,相比前作DCVC-HEM,实现了约23.5%的码率节省,证明了在更广阔的时空域内进行上下文建模的有效性。针对单一模型难以覆盖广泛码率范围的难题,Li等[69]提出了DCVC-FM。该模型采用特征调制方案,设计了可学习的量化缩放器和均匀量化参数采样机制。结合周期性的帧内刷新策略,DCVC-FM能够通过单一模型支持极宽的质量范围,其性能超过H.266/VVC达25.5%的码率节省,展现了极强的时域稳定性。而面向工业界的实时应用需求,最新的DCVC-RT[70]采用效率驱动的设计理念,该模型摒弃了复杂的显式光流运动估计模块实现了实时编码,转而采用“隐式时域建模”策略,并利用单一低分辨率潜在表示替代传统的渐进式下采样。该方案在NVIDIA A100上实现了1 080P视频高达125.2 fps的编码速度。在保持实时性的前提下,其压缩性能仍优于H.266/VVC达21%,解决了高压缩率与实时解码难以兼得的行业瓶颈。

在其他条件编码框架的相关研究中,叶枫等[71]提出了基于分层双向参考结构的端到端视频编码框架,通过设计参数共享的运动编解码器、双向缩放运动先验及可信运动建模技术,有效提升了无人机视频压缩的率失真性能,其压缩性能优于传统H.266/VVC标准。Sheng等[72]针对现有时域上下文挖掘框架和处理局部运动不一致性与物体遮挡方面的局限,提出了基于空间分解与时间融合的帧间预测方法,其设计了结构与细节分量,通过将视频帧解耦为低频结构分量与高频细节分量并独立进行运动建模,有效解决了前景与背景不一致问题。为了解决通道间冗余干扰与特征表示优化不足的问题,Guo等[73]提出了增强上下文挖掘与滤波网络,通过增强上下文挖掘模块去除通道间冗余,并设计了基于Transformer的环路滤波机制,利用全分辨率转置注意力捕捉全局相关性,从

而在优化特征表示的同时有效提升了重建质量与压缩性能。相比DCVC在PSNR指标上实现了9.85%和在MS-SSIM指标上12.19%的码率节省，且在MS-SSIM指标上超越了VVC（LDP配置）平均6.7%。Wu等[74]针对条件编码框架种误差传播难以抑制的问题，提出了ConFRE框架，重点研究了帧内循环上下文滤波。该方法通过在编码循环内对时域上下文进行精细化滤波来抑制误差传播，并结合自适应编码决策策略动态开启滤波，同时配合环外重构增强模块，在条件编码基准上进一步实现了7.71%的码率节省。Wang等[75]设计了一个多尺度运动感知（MASTC-VC）模块以实现由粗到细的时空运动特征增强，并引入时空通道上下文模块配合非均匀通道分组策略，充分挖掘了潜在表征相关性，减少了码率消耗。此外，Zhang等[76]提出的FLAVC进一步在特征层级引入了注意力机制，根据运动幅度自适应加权不同的特征区域，有效解决了复杂运动场景下的特征错位问题，相比H.265/HEVC实现了高达61.46%的码率增益。针对仅依赖时域参考的条件编码在处理大幅运动与新兴物体的局限，Bian等[77]提出了空间嵌入式视频编码框架（SEVC），其嵌入了一个低分辨率基础编解码器以提供空间参考，通过运动与特征协同增强模块生成混合时空上下文，并利用空间引导的潜在先验策略有效对齐了多帧时域特征。这一设计有效解决了传统神经视频编码器在复杂场景下的对齐难题，相比于DCVC-DC和DCVC-FM等SOTA方法，SEVC实现了约11.9%的额外码率节省。为了克服现有深度编解码器在码率与复杂度调节上的灵活性缺失，Wei等[78]提出了可调节深度视频编解码器（RDVC），通过设计自适应特征压缩网络，由粗到细的两阶段调制实现码率的平滑调整，并利用时空特征传播机制稳定参考质量，借助可瘦身卷积技术实现了在单一权重下对解码复杂度的动态控制，RDVC在提供极高灵活性的同时并未牺牲性能，其在MS-SSIM指标上相比H.266/VVC平均减少了58.12%的比特数，在PSNR指标上也实现了9.35%的节省，展示了深度视频编码在实用化方向的巨大潜力。

3.3 基于生成式模型的视频编码

尽管上述基于条件编码的视频压缩通过引入时空上下文显著降低了帧间冗余并提升了率失真性能，但在极低码率场景下，这类基于确定性映射的方法往往会产生严重的模糊效应，难以满足人类视觉对高感知质量的需求。为了突破这一瓶颈，研究者开始探索将生成式模型引入压缩框架，利用其强大的先验分布建模能力来合成高频细节。Mentzer等[79]率先提出了首个基于生成对抗网络的神经视频压缩范式。该框架确立了“细节合成与时域传播”的核心策略，通过在I帧中合成高度逼真的纹理细节，并结合高质量光流网络与解耦缩放空间变换机制，将已生成的细节精确传播至后续P帧。这一设计有效解决了传统神经压缩在低码率下普遍存在的画面模糊问题。在此基础上，为了进一步增强视频序列的时空连贯性并抑制生成式方法常见的闪烁伪影，Yang等[80]提出了感知学习视频压缩框架。该方法采用循环自动编码器作为生成器，并创新性地设计了一个循环条件判别器，通过综合评估潜表现、时域运动信息以及循环单元的隐藏状态，强制生成器在空间逼真度与时域一致性之间达成平衡。实验证明，该框架在超低码率下的感知质量显著优于H.265/HEVC。随着深度学习架构的演进，研究者开始探索更高效的特征提取与建模方式。Du等[81]提出了结合Transformer的上下文生成视频压缩模型（CGVC-T）。该模型通过混合Transformer-convolution结构捕捉视频帧内的全局与局部相关性，并首次将上下文编码机制引入生成式框架。相比于传统的预测残差编码，该方法直接利用前序帧提供的丰富上下文特征引导目标帧的重构，显著提升了编码效率。在HEVC和UVG等数据集上，其感知指标全面超越了现有的神经编解码器及H.266/VVC标准。而在追求极致压缩率的场景中，Li等[82]进一步挖掘了预训练扩散模型的预测潜能。该方案在解码端利用条件扩散模型递归地预测后续帧内容，仅在重构质量低于预设阈值时才对新帧进行编码以重启预测。这种基于“预测-修正”的极端压缩策略在比特率低至0.02 Bpp以下时，仍能维持极

高的感知清晰度与时空相干性。针对扩散模型在视频任务中面临的推理延迟与码率适配难题,Ma等[83]提出了基于扩散模型的感知视频压缩框架DiffVC。该框架将基础扩散模型作为生成式解码器嵌入条件编码架构中,利用其强大的生成先验并结合时域上下文引导高质量重建。为了克服扩散模型采样过程冗长的问题,作者提出了时域扩散信息复用策略以加速推理过程,并设计了基于量化参数的提示机制来调制特征生成。该方案在仅损失约1.96%感知性能的前提下,将推理速度提升了47%,并实现了稳健的可变比特率控制。Mall等[84]在CRAM框架中探索了自适应刷新编码在长视频持续学习中的应用。虽然其核心目标是持续学习,但其利用在线训练的视频压缩器对神经代码进行不断的压缩与解吸,通过"缓冲刷新"策略缓解了压缩模型在处理流式数据时的灾难性遗忘,展示了生成式逻辑在极低带宽下维持长时视频表征的能力。最后,Liu等[85]提出了更为先进的统一压缩框架I^{2}VC。该框架通过时空可变速率编解码器与隐式帧间特征对齐模块,利用去噪扩散隐式模型逆转技术在特征空间直接实现帧间一致性对齐,彻底摆脱了对显式运动估计与补偿的依赖。这一高度集成的框架能够统一处理AI、LD及RA等多种视频配置,其感知重建性能相比H.266/VVC实现了58.4%的巨大增益。

3.4 基于隐式神经表达的视频编码

为了突破传统视频表示将视频视为帧序列的局限,基于隐式神经表达(Implicit neural representation,INR)的视频编码方法也颇受关注。INR采用"视频即函数"的范式,将离散的视频数据建模为定义在时空域上的连续映射,并通过优化神经网络的权重参数来实现对视频内容的隐式编码。

Chen等[86]率先提出了视频神经表达(NeRV),将传统的像素级映射革新为帧级映射。该框架通过一个以时间坐标为输入的层级卷积网络,直接输出全分辨率图像,并通过模型剪枝与量化技术实现压缩,在解码速度上相较于传统的像素级INR提升了数十倍。为了解决NeRV在处理高分辨率视频时参数分布不均及收敛慢的问题,Li等[87]提出了E-NeRV,通过将时空上下文进行解耦建模,利用固定的空间网格坐标与可学习的时间索引分别表征视频特征,在显著减少冗余参数的同时,实现了8倍以上的收敛加速。Bai等[88]则从局部相关性出发,提出了分块神经表示模型PS-NeRV。该模型通过将视频分解为空间重叠的补丁,并引入自适应实例归一化机制来调制卷积层特征,使网络能够更精准地拟合高频纹理,有效克服了全图重构在极低码率下画面过平滑的缺陷。随着对表征精度要求的进一步提升,Chen等[89]在HNeRV中引入了内容自适应机制,构建了首个混合神经表示框架。该框架通过增加一个轻量化编码器,根据输入视频生成动态的、内容相关的嵌入向量作为解码器输入,并重新设计了解码模块以实现参数在层间的均匀分布。这一改进使模型在视频重归任务中提升了约4.7 dB的PSNR,展现出极强的内部泛化能力。针对时域冗余利用不足的问题,Lee等[90]提出了流引导的帧级表示模型FFNeRV。该框架创新性地结合了传统视频编解码器的运动建模思想,通过多尺度时域网格提取空间特征,并预测光流图与聚合权重,利用扭曲机制重构相邻帧之间的共性像素,极大地增强了模型处理复杂动态场景的能力,在UVG数据集上达到了与H.264/AVC相当的压缩效率。针对现有INR方法无法灵活调节码率、需为不同质量等级重复训练独立模型的局限,Wu等[91]提出了质量可伸缩隐式神经表示模型QS-NeRV。该模型采用层级化设计,由基础层和可扩展增强层组成,并引入可逆跳跃连接,在不增加额外比特开销的情况下实现编解码间的高效信息交换。该框架支持单一权重下的实时(>30 fps)质量可伸缩解码,在保持轻量化解码优势的同时,显著提升了视频重构与插值的精度。最近,Kwan等[92]提出了更为先进的分层编码架构HiNeRV,该模型通过结合分层位置编码与深度宽网络结构,显著提升了INR的特征表征容量。HiNeRV不仅实现了帧级与块级表示的统一,使其在硬件部署上更具灵活性,还设计

了精细化的剪枝与量化感知训练流水线。实验结果表明，HiNeRV在性能上实现了跨越式发展，在UVG数据集上相比HNeRV实现了72.3%的比特率节省，并成为首个在PSNR指标上全面超越H.265/HEVC及DCVC等端到端神经编解码器的隐式表示方法。Ling等[93]针对多视图视频提出了MV-MGINR框架，采用了多重网格隐式表达。该方法通过结合时间索引网格、视角索引网格以及时空集成网格，分别捕捉多视图视频中的共性特征与局部细节。通过这种分层网格设计配合运动感知损失函数，该方法有效平抑了INR在处理大范围运动时的伪影难题，相比MPEG的TMIV模型实现了高达72.3%的码率节省。Gao等[94]提出了生成式隐式视频压缩框架GIViC，其创新性地将隐式扩散过程与Transformer架构结合，利用层级门控线性注意力捕捉长程时空依赖，实现了由粗到精的扩散采样。实验证明，GIViC是首个在随机访问配置下率失真性能全面超越H.266/VVC的INR编解码器，在UVG数据集上实现了15.94%的码率节省。

为了直观地展示端到端神经视频编码近年来的快速发展，图6汇集了十余种神经编解码器与标准H.266/VVC及H.265/HEVC在UVG 1 080P数据集上的客观性能对比。从图中可以看出，2025年的最新SOTA模型已在全码率范围内实现了对H.266/VVC的稳健压制。这一演进趋势有力地证明了深度神经编码架构在挖掘时空冗余方面的巨大潜力，标志着其已正式进入全面竞争甚至引领行业标准的阶段。

为了进一步说明端到端视频编码的发展，表6系统梳理了4类主流神经视频编码模型（残差编码、条件编码、生成式模型、隐式神经表达）的核心信息。早期的残差编码（如DVC、M-LVC）主要解决端到端框架的可行性与基础性能；随着技术演进，条件编码（如DCVC系列）凭借上下文先验的引入，逐渐在码率节省与实时性上占据主导地位；而生成式模型聚焦低码率感知质量优化，从缓解画面模糊到DiffVC平衡速度与感知损失，逐步完善实用化能力；隐式神经表达类模型则凭借帧级隐式表示在比特率压缩上表现突出，HiNeRV较传统INR实现72.3%的比特率节省，GIViC则在随机访问场景下超越H.266/VVC标准。可以看出，近年来，端到端视频压缩正在从对标传统标准向追求极致压缩与计算效率转化。

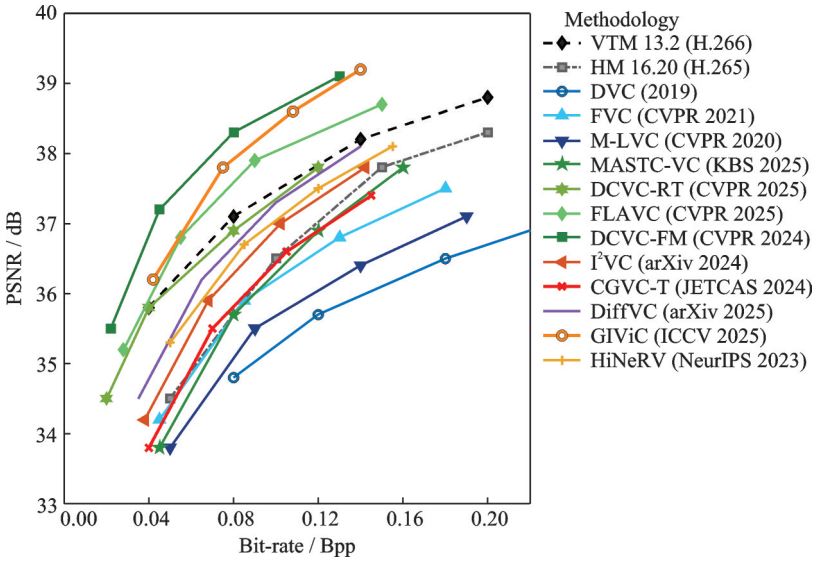

图6 UVG数据集上的R-D性能对比

Fig.6 Rate-distortion (R-D) performance comparison on the UVG dataset

表6 基于深度学习的端到端视频编码技术比较

Table 6 Comparison of end-to-end video coding techniques based on deep learning

模型	发表年份	类型	核心创新点	关键性能指标
DVC[52]	2019	残差编码	首个端到端深度学习视频压缩框架,用CNN替换传统编码所有模块,端到端率失真优化	PSNR优于H.264/AVC,MS-SSIM与H.265/HEVC持平
M-LVC[54]	2020	残差编码	多帧预测机制,联合多相邻帧与运动信息,抑制预测误差累积	PSNR在低码率段领先H.265/HEVC
FVC[56]	2021	残差编码	在特征空间完成运动补偿、残差计算与压缩,引入可变形卷积增强对齐能力	较H.265/HEVC实现14.18%~28.71%码率节省,UVG数据集码率增益达28.71%
MMVC[60]	2023	残差编码	特征域多预测模式设计,自适应选择最优运动描述方式	率失真性能可与H.266/VVC抗衡
DCVC[65]	2021	条件编码	引入条件编码框架,以特征域上下文作为先验指导概率建模与重建	1 080P数据集较H265/HEVC码率节省26.0%
DCVC-TCM[66]	2022	条件编码	多尺度时域上下文挖掘,摒弃自回归空间先验以提升推理效率	较H.265/HEVC码率节省14.4%,MS-SSIM超H.266/VVC 21.1%
DCVC-HEM[67]	2022	条件编码	混合时空熵模型捕捉时空统计相关性,引入多粒度量化	较H.266/VVC最高配置码率节省18.2%
DCVC-DC[68]	2023	条件编码	多样化上下文建模策略,结合四叉树结构实现高效熵编码	超越ECM,较DCVC-HEM码率节省23.5%
DCVC-FM[69]	2024	条件编码	特征调制与可学习量化缩放器设计,支持周期性帧内刷新	性能超H.266/VVC 25.5%,覆盖宽质量区间
DCVC-RT[70]	2025	条件编码	隐式时域建模与低分辨率潜在表示,完全摒弃显式光流	1 080P编码速度达125.2 fps(A100),性能超H.266/VVC 21%
FLAVC[76]	2025	条件编码	特征层级注意力机制,根据运动幅度自适应加权特征区域	较H.265/HEVC码率增益61.46%
SEVC[77]	2025	条件编码	空间嵌入式编码框架,低分辨率基础编解码器提供空间参考	较DCVC-DC/DCVC-FM额外码率节省11.9%
RDVC[78]	2025	条件编码	自适应特征压缩网络,两阶段特征调制与轻量化卷积设计	MS-SSIM下比特数减少58.12%,PSNR码率节省9.35%
CGVC-T[79]	2022	生成式模型	Transformer与卷积混合结构,引入上下文编码的生成式框架	感知指标超越神经编解码器及H.266/VVC
DiffVC[83]	2025	生成式模型	扩散模型作为生成式解码器,复用时域扩散信息加速推理	推理速度提升47%,感知性能损失仅1.96%
NeRV[86]	2021	隐式神经表达	帧级隐式神经表示,层级卷积网络直接生成全分辨率帧	解码速度较传统INR提升数十倍
E-NeRV[87]	2022	隐式神经表达	时空上下文解耦建模,空间坐标与时间索引联合表示	参数冗余显著减少,收敛速度提升8倍以上。
HiNeRV[92]	2023	隐式神经表达	分层位置编码与宽深网络结构,结合感知剪枝与量化训练	较HNeRV比特率节省72.3%,PSNR在高码率段超越H265/HEVC与DCVC
GIViC[94]	2025	隐式神经表达	隐式扩散与Transformer融合,层级门控线性注意力机制	随机访问配置下超越H.266/VVC,UVG码率节省15.94%

尽管上述的视频压缩技术在视觉质量上取得了突破性进展，但它们主要仍以服务“人类视觉系统”为单一目标。然而，随着人工智能和机器学习等领域的不断发展，各种类型的数据来源现今都可以通过神经网络进行高效处理，据思科统计，视频数据约占八成以上的比例都用于机器智能分析及各种算法处理。传统的以像素重建为目标的编码方式，对于目标检测、语义分割等机器视觉任务而言，往往存在冗余或关键特征丢失。因此，未来的视频编码正逐渐向人机协同的方向演进，其旨在探索可分级或特征层面的压缩表示，使得同一码流既能被解码为供人观看的高质量视频，又能直接被机器提取特征用于下游任务。

端到端视频编码作为更彻底的技术革新，其同样面临一系列亟待解决的瓶颈。首先，端到端网络本质上是黑盒模型，其比特分配机制与率失真优化缺乏传统编码器所具有的理论解释性和精细调控能力，这给实际系统的调试、优化和码率控制带来了困难。其次，端到端模型的性能高度依赖训练数据的分布，在面向训练集中未出现的内容（如剧烈运动、复杂纹理）时，其泛化能力可能出现显著下降，导致实际应用中的性能不稳定。再者，当前端到端编解码器通常依赖GPU等高性能硬件加速，计算复杂度远超传统编解码器，在移动端、嵌入式等资源受限场景的部署面临严峻挑战。并且，端到端架构与传统编码标准体系不兼容，难以融入现有的视频传输、存储与播放生态，产业化落地路径尚不明朗。因此，提升端到端编码的可解释性、泛化能力和生态兼容性，将是推动其从学术研究走向实用化的核心攻关方向。

4 总结与展望

随着互联网视频业务向超高清、沉浸式方向的爆发式增长，海量数据对传输带宽与存储空间提出了严峻挑战，视频编码技术因此成为解决这一瓶颈的关键。纵观当前技术发展格局，传统视频编码技术仍占据当前应用的主导地位。尽管基于人工设计的传统混合编码框架因模块独立优化而逐渐触及性能天花板，但凭借其成熟的标准化体系、完善的硬件生态以及高度的可实现性，传统视频编码标准（如H.265/HEVC、H.266/VVC）在未来相当长一段时间内仍将是工业界实际落地的主流解决方案。其次，结合深度学习的视频编码技术将作为一种务实的增强手段长期存在。这类方法通过将神经网络嵌入传统框架的关键环节来提升自适应能力，虽然未能摆脱传统架构的束缚，导致其性能提升空间受限，但正因为其保留了传统框架的骨干结构，相比于完全颠覆性的端到端方案，该类技术在工程化部署与实际应用落地方面具有更快的速度和更高的可行性，因此仍是学术界与工业界持续深耕的重要方向。最后，基于深度学习的端到端视频编码框架代表了视频压缩领域的未来演进方向。该框架通过全链路联合优化，打破了传统人工设计参数的依赖，在率失真性能上已展现出超越传统视频编码标准的巨大潜力。尽管目前端到端技术仍面临计算复杂度高、泛化能力待增强等现实问题，但其从根本上重塑了编码范式，随着算法迭代与算力提升，有望成为下一代视频压缩技术的主流。基于上述分析，下面对深度学习驱动的视频编码技术的未来发展趋势进行展望：

（1）标准化发展

目前，工业界和学术界也在致力于研发超越VVC标准的下一代视频编码技术，而基于深度学习的视频编码技术凭借其卓越的压缩潜力和有效算法，已成为下一代标准的核心焦点。特别是人工智能运动图像、音频和数据编码（Moving picture, audio and data coding by artificial intelligence, MPAI）工作组的最新动向，清晰地描述了该技术的发展路线，其同步推动增强视频编码（Enhanced video coding, EVC）和端到端视频编码（End to end video coding, EEV）两个研究项目，其中，MPAI-EVC旨在利用神经网络替代或改进现有工具，以满足中短期的视频编码需求；而MPAI-EEV作为更彻底的视频编码技术革新，致力于构建全新的完全基于深度学习网络的独立标准，以应对长期的视频编码挑战[95-96]。这种“双轨并

行”的战略布局既务实又具有前瞻性,不仅印证了深度学习技术的有效性,更凸显了构建高效、独立的端到端视频编码标准将是未来视频编码技术演进的核心攻坚方向。在标准化发展方面,我们强调构建具备高度可扩展性的架构设计思路,提出应探索支持传统工具与深度学习模块灵活融合的标准化机制,并研究如何建立能够适应端到端模型快速演进的迭代更新框架,以突破传统标准修订周期长、适应性不足的局限。

(2) 计算复杂度和实时性

对于视频编码系统而言,压缩效率和视觉质量至关重要,但其计算复杂度也是不容忽视的重要因素。尽管基于深度学习的视频编码方法在编码性能方面已优于传统编码方法,但此类方法通常需要大量计算资源和存储空间以处理日益增长的网络参数规模,从而导致编解码时间过长,难以满足实时应用需求[97]。此外,这类方法主要依赖于GPU加速,对于无法访问显卡资源的终端设备(如手机、电视等)存在较大限制。近年来,基于深度学习的编码器性能的提升无不伴随着计算复杂度的提高,阻碍了在移动平台、应用软件等资源受限环境中的普及利用。因此,基于深度学习的视频编码技术的发展趋势不仅要进一步关注压缩性能,还应重点解决计算复杂度和实时性等问题。未来研究可深度挖掘于深度网络模型的潜力,探索轻量化的网络模型设计,以及通过利用知识蒸馏、剪枝和量化等操作来减少模型参数量,并且还可以开发专用的网络加速器,或结合电子产品中的硬件DNN加速器来优化时间效率问题,以推动基于深度学习的视频编码网络模型的应用。

(3) 多样化视频内容的适应性

未来伴随着数据视频应用场景的日益扩展和需求层次的不断提升,视频编码技术不仅应在标准数据集上的压缩性能表现出色,更需要在面对多样化、极端化的实际场景时表现出足够的泛化能力与鲁棒性[26,98]。现有方法在处理剧烈运动、复杂纹理等复杂非典型场景时,其性能往往出现明显下降,限制了其在真实开放环境中的适用性。因此,构建一个具备强泛化能力和多场景适应性的通用化视频编码框架也是未来需重点突破的方向。然而,现有的基于深度学习的视频编码算法多针对固定码率或特定任务进行编码优化,通常为不同码率训练多个独立模型,这不仅增加了存储和部署开销,也导致其码率-失真特性呈现离散且不连续的特点。因此,现有方法难以实现动态场景下码率变化需求,缺乏对码率和率失真优化的有效控制。要推动基于深度学习的视频编码的应用,必须设计连续可调的率失真优化机制,发展能够根据场景特征、码率需求与智能分析任务进行自适应动态调整的灵活码率控制策略,从而扩展其在多样化视频内容与动态条件下的适用范围。最后,随着多媒体内容的多样化发展,面向图像、视频、文本和音频等多模态数据的协同编码与联合压缩也是未来智能编码的方向,不仅有助于进一步提升压缩效率,也将为未来媒体传输与智能视觉分析提供更高效的数据表示,以满足用户和机器任务的多样化需求。

(4) 面向机器视觉的视频编码

随着智慧城市、智慧交通及智能监控等智能化社会应用的迅猛发展,面向机器视觉的编码需求急剧上升,视频数据已从传统意义上为人眼观看的媒介,逐渐转变为机器分析和理解的关键信息源[99]。在此背景下,机器视觉的重要性日益凸显,其作为智能化系统的“眼睛”,承担着对海量视频进行自动解析、目标识别、行为分析和决策支持等核心任务。因此,对数字视频的要求不仅需要满足高压缩率以减少数据量或满足人眼主观感知质量,更在于保证视频数据具备高度的机器可理解性,以便更好地服务于机器任务。并且,相关研究机构也对机器视觉编码技术开展了标准化研究,如MPEG(Moving picture experts group)成立VCM(Video coding for machines)工作组探索面向人机协同的视频编码标准[100]。因此,面向人眼感知和机器视觉性能的“人机协同”的视频编码方法也是未来发展的重要方向之一。未来视频编码技术研究不应只将“保真度”局限于像素级的人眼视觉重建,还应深入探索对机器分

析至关重要的语义特征和结构信息，从而持续推进智能化任务的发展。在面向机器视觉的视频编码方面，我们进一步阐述了“人机协同”率失真优化理论框架的研究设想。该框架应能够显式建模机器视觉任务性能（如目标检测准确率、识别精度等）在编码过程中的损失，并将其纳入优化目标函数，在码率、人眼感知质量与机器可理解性之间建立联合优化机制，从而实现像素级保真度与语义级保真度之间的最优平衡。在此基础，探索同时满足面向人眼视觉与机器视觉的新的评价体系也是重要的研究方向，从而无需在面向机器任务时使用mAP、MOTA等评价指标，面向人眼视觉时使用PSNR、MS-SSIM及BD-rate等指标。

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] BROSS B, WANG Y K, YE Y, et al. Overview of the versatile video coding (VVC) standard and its applications[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(10): 3736-3764.

[4] LIN L, YU S, ZHOU L, et al. PEA265: Perceptual assessment of video compression artifacts[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(11): 3898-3910.

[5] BALLÉ J, LAPARRA V, SIMONCELLI E P. End-to-end optimized image compression[EB/OL]. (2017-03-03). https://arxiv.org/abs/1611.01704.

[6] 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. [S.l.]: IEEE, 2019: 221-231.

[7] BALLÉ J, MINNEN D, SINGH S, et al. Variational image compression with a scale hyperprior[EB/OL]. (2018-02-01). https://arxiv.org/abs/1802.01436.

[8] TODERICI G, O'MALLEY S M, HWANG S J, et al. Variable rate image compression with recurrent neural networks[EB/OL]. (2016-03-01). https://arxiv.org/abs/1511.06085.

[9] ZHANG X, WANG S, GU K, et al. Just-noticeable difference-based perceptual optimization for JPEG compression[J]. IEEE Signal Processing Letters, 2016, 24(1): 96-100.

[10] LI M, ZUO W, GU S, et al. Learning convolutional networks for content-weighted image compression[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 3214-3223.

[11] MENTZER F, AGUSTSSON E, TSCHANNEN M, et al. Conditional probability models for deep image compression[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 4394-4402.

[12] 岳爽, 陈喆, 殷福亮. 极低比特率图像压缩技术综述[J]. 数据采集与处理, 2025, 40(1): 102-116.
YUE Shuang, CHEN Zhe, YIN Fuliang. A review of very low bit-rate image compression techniques[J]. Journal of Data Acquisition and Processing, 2025, 40(1): 102-116.

[13] WALLACE G K. The JPEG still picture compression standard[J]. IEEE Transactions on Consumer Electronics, 2002. DOI: 10.1109/30.12.5072.

[14] SKODRAS A, CHRISTOPOULOS C, EBRAHIMI T. The JPEG 2000 still image compression standard[J]. IEEE Signal Processing Magazine, 2002, 18(5): 36-58.

[15] BELLARD F. BPG image format[EB/OL]. (2018-10-30), https://bellard.org/bpg/.

[16] ALSHINA E, ASCENSO J, EBRAHIMI T. JPEG AI: The first international standard for image coding based on an end-to-end learning-based approach[J]. IEEE MultiMedia, 2024, 31(4): 60-69.

[17] LIU D, LI Y, LIN J, et al. Deep learning-based video coding: A review and a case study[J]. ACM Computing Surveys (CSUR), 2020, 53(1): 11.

[18] LIU Z, YU X, GAO Y, et al. CU partition mode decision for HEVC hardwired intra encoder using convolution neural network

[J]. IEEE Transactions on Image Processing, 2016, 25(11): 5088-5103.

[19] SONG R, LIU D, LI H, et al. Neural network-based arithmetic coding of intra prediction modes in HEVC[C]//Proceedings of 2017 IEEE Visual Communications and Image Processing (VCIP). [S.l.]: IEEE, 2017: 1-4.

[20] ZHAO L, WANG S, ZHANG X, et al. Enhanced CTU-level inter prediction with deep frame rate up-conversion for high efficiency video coding[C]//Proceedings of 2018 25th IEEE International Conference on Image Processing (ICIP). [S.l.]: IEEE, 2018: 206-210.

[21] LU G, OUYANG W, XU D, et al. Deep Kalman filtering network for video compression artifact reduction[C]//Proceedings of the European Conference on Computer Vision (ECCV). [S.l.]: Springer, 2018: 591-608.

[22] YANG R, XU M, WANG Z, et al. Multi-frame quality enhancement for compressed video[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 6664-6673.

[23] ZHAO Z, WANG S, WANG S, et al. CNN-based bi-directional motion compensation for high efficiency video coding[C]//Proceedings of 2018 IEEE International Symposium on Circuits and Systems (ISCAS). [S.l.]: IEEE, 2018: 1-4.

[24] 王婷, 何小海, 孙伟恒, 等.结合卷积神经网络的HEVC帧内编码压缩改进算法[J]. 太赫兹科学与电子信息学报, 2020, 18(2): 291-297.
WANG Ting, HE Xiaohai, SUN Weiheng, et al. Improved HEVC intra coding compression algorithm combined with convolutional neural network[J]. Journal of Terahertz Science and Electronic Information Technology, 2020, 18(2): 291-297.

[25] HOANG T M, ZHOU J. Recent trending on learning based video compression: A survey[J]. Cognitive Robotics, 2021, 1: 145-158.

[26] MA S, ZHANG X, JIA C, et al. Image and video compression with neural networks: A review[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(6): 1683-1698.

[27] 高文, 赵德斌, 马思伟. 数字视频编码技术原理[M]. 第二版. 北京: 科学出版社, 2021.
GAO Wen, ZHAO Debin, MA Siwei. Principles of digital video coding technology [M]. 2nd ed. Beijing: Science Press, 2021.

[28] 何小海. 数字图像通信[M]. 成都: 四川大学出版社, 2010.
HE Xiaohai. Digital image communication[M]. Chengdu: Sichuan University Press, 2010.

[29] 裴智勇, 张春红. H.26x与MPEG-x[J]. 电信工程技术与标准化, 2005, 18(2): 32-36.
PEI Zhiyong, ZHANG Chunhong. H.26x and MPEG-x[J]. Telecom Engineering Technics and Standardization, 2005, 18(2): 32-36.

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

[31] DUMAS T, ROUMY A, GUILLEMOT C. Context-adaptive neural network-based prediction for image compression[J]. IEEE Transactions on Image Processing, 2019, 29: 679-693.

[32] JIN Z, AN P, SHEN L. Video intra prediction using convolutional encoder decoder network[J]. Neurocomputing, 2020, 394: 168-177.

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

[34] WANG Y, FAN X, LIU S, et al. Multi-scale convolutional neural network-based intra prediction for video coding[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(7): 1803-1815.

[35] ZHU L, KWONG S, ZHANG Y, et al. Generative adversarial network-based intra prediction for video coding[J]. IEEE Transactions on Multimedia, 2019, 22(1): 45-58.

[36] HUO S, LIU D, WU F, et al. Convolutional neural network-based motion compensation refinement for video coding[C]//Proceedings of 2018 IEEE International Symposium on Circuits and Systems (ISCAS). [S.l.]: IEEE, 2018: 1-4.

[37] WANG Y, FAN X, XIONG R, et al. Neural network-based enhancement to inter prediction for video coding[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(2): 826-838.

[38] YAN N, LIU D, LI H, et al. Convolutional neural network-based fractional-pixel motion compensation[J]. IEEE Transactions

on Circuits and Systems for Video Technology, 2018, 29(3): 840-853.

[39] LIN J, LIU D, LI H, et al. Generative adversarial network-based frame extrapolation for video coding[C]//Proceedings of 2018 IEEE Visual Communications and Image Processing (VCIP). [S.l.]: IEEE, 2018: 1-4.

[40] XIONG L, ZHOU W, ZHOU X, et al. Saliency aware fast intra coding algorithm for HEVC[C]//Proceedings of 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). [S.l.]: IEEE, 2016: 1-5.

[41] JIANG X, SONG T, ZHU D, et al. Quality-oriented perceptual HEVC based on the spatiotemporal saliency detection model [J]. Entropy, 2019, 21(2): 165.

[42] WU Y, WANG Z, CHEN W, et al. Perceptual VVC quantization refinement with ensemble learning[J]. Displays, 2021, 70: 102103.

[43] ZHU S, CHANG Q, LI Q. Video saliency aware intelligent HD video compression with the improvement of visual quality and the reduction of coding complexity[J]. Neural Computing and Applications, 2022, 34(10): 7955-7974.

[44] LI W, JIANG X, JIN J, et al. Saliency-enabled coding unit partitioning and quantization control for versatile video coding[J]. Information, 2022, 13(8): 394.

[45] HUANG Z, GUO X, SHANG M, et al. An efficient QP variable convolutional neural network based in-loop filter for intra coding[C]//Proceedings of 2021 Data Compression Conference (DCC). [S.l.]: IEEE, 2021: 33-42.

[46] ZHANG H, LIU Y, JUNG C, et al. RTNN: A neural network-based in-loop filter in VVC using resblock and transformer[J]. IEEE Access, 2024, 12: 104599-104610.

[47] LIU Y, JUNG C. DQT-CALF: Content adaptive neural network based In-Loop filter in VVC using dual query transformer[J]. Neurocomputing, 2025, 637: 130064.

[48] ZHANG T, TENG Q, HE X, et al. Multi-scale inter-communication spatio-temporal network for video compression artifacts reduction[J]. IEEE Transactions on Circuits and Systems Ⅱ: Express Briefs, 2022, 70(3): 1229-1233.

[49] WANG J, DENG X, XU M, et al. Multi-level wavelet-based generative adversarial network for perceptual quality enhancement of compressed video[C]//Proceedings of European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 405-421.

[50] 于海, 杨磊, 高阳, 等. 基于块编码特点的压缩视频质量增强算法[J]. 北京工业大学学报, 2024, 50(9): 1069-1076.
YU Hai, YANG Lei, GAO Yang, et al. Compressed video quality enhancement algorithm based on block coding characteristics[J]. Journal of Beijing University of Technology, 2024, 50(9): 1069-1076.

[51] ZHU Q, HAO J, DING Y, et al. CPGA: Coding priors-guided aggregation network for compressed video quality enhancement [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2024: 2964-2974.

[52] LU G, OUYANG W, XU D, et al. DVC: An end-to-end deep video compression framework[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 11006-11015.

[53] 王惠之, 王潇毅, 谢冠超, 等. 内容自适应驱动的深度视频压缩算法[J/OL]. 重庆邮电大学学报(自然科学版), 2026: 1-10 [2026-03-14]. https://link. cnki.net/urlid/50.1181.N.20251125.1747.026.
WANG Huizhi, WANG Xiaoyi, XIE Guanchao, et al. Content-adaptive driven deep video compression algorithm[J/OL]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2026: 1-10[2026-03-14]. https://link. cnki.net/urlid/50.1181.N.20251125.1747.026.

[54] LIN J, LIU D, LI H, et al. M-LVC: Multiple frames prediction for learned video compression[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2020: 3546-3554.

[55] 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. [S.l.]: IEEE, 2020: 8503-8512.

[56] HU Z, LU G, XU D. FVC: A new framework towards deep video compression in feature space[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2021: 1502-1511.

[57] 徐智慧, 毕晓东, 杨红, 等. 基于可变形帧和自回归通道预测的视频压缩算法[J]. 通信技术, 2025, 58(3): 262-269.
XU Zhihui, BI Xiaodong, YANG Hong, et al. Video compression algorithm based on deformable frames and autoregressive channel prediction[J]. Communications Technology, 2025, 58(3): 262-269.

[58] 张秋建, 熊淑华, 杨红, 等. 基于运动先验引导的端到端视频压缩算法[J]. 无线电工程, 2025, 55(12): 2452-2460.
ZHANG Qiujian, XIONG Shuhua, YANG Hong, et al. An end-to-end video compression algorithm guided by motion priors [J]. Radio Engineering, 2025, 55(12): 2452-2460.

[59] GAO H, CUI J, YE M, et al. Structure-preserving motion estimation for learned video compression[C]//Proceedings of the 30th ACM International Conference on Multimedia. [S.l.]: ACM, 2022: 3055-3063.

[60] LIU B, CHEN Y, MACHINENI R C, et al. MMVC: Learned multi-mode video compression with block-based prediction mode selection and density-adaptive entropy coding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2023: 18487-18496.

[61] WANG H, CHEN Z, CHEN C W. Learned video compression via heterogeneous deformable compensation network[J]. IEEE Transactions on Multimedia, 2023, 26: 1855-1866.

[62] ZHANG T, HE X, TENG Q, et al. Spatio-temporal adaptive weighted fusion network for compressed video quality enhancement[J]. IEEE Transactions on Circuits and Systems Ⅱ: Express Briefs, 2024, 71(12): 5064-5068.

[63] 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. [S.l.]: IEEE, 2020: 7939-7948.

[64] RIPPEL O, ANDERSON A G, TATWAWADI K, et al. ELF-VC: Efficient learned flexible-rate video coding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2021: 14479-14488.

[65] LI J, LI B, LU Y. Deep contextual video compression[J]. Advances in Neural Information Processing Systems, 2021, 34: 18114-18125.

[66] SHENG X, LI J, LI B, et al. Temporal context mining for learned video compression[J]. IEEE Transactions on Multimedia, 2022, 25: 7311-7322.

[67] LI J, LI B, LU Y. Hybrid spatial-temporal entropy modelling for neural video compression[C]//Proceedings of the 30th ACM International Conference on Multimedia. [S.l.]: ACM, 2022: 1503-1511.

[68] LI J, LI B, LU Y. Neural video compression with diverse contexts[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2023: 22616-22626.

[69] LI J, LI B, LU Y. Neural video compression with feature modulation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2024: 26099-26108.

[70] JIA Z, LI B, LI J, et al. Towards practical real-time neural video compression[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, TN, USA: IEEE, 2025: 12543-12552.

[71] 叶枫, 董凡可, 贾川民. 端到端智能视频压缩技术及其在无人机中的应用[J]. 数据采集与处理, 2025, 40(2): 303-319.
YE Feng, DONG Fanke, JIA Chuanmin. End-to-end intelligent video compression technology and its application in unmanned aerial vehicles[J]. Journal of Data Acquisition and Processing, 2025, 40(2): 303-319.

[72] SHENG X, LI L, LIU D, et al. Spatial decomposition and temporal fusion based inter prediction for learned video compression [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(7): 6460-6473.

[73] GUO H, KWONG S, YE D, et al. Enhanced context mining and filtering for learned video compression[J]. IEEE Transactions on Multimedia, 2023, 26: 3814-3826.

[74] WU Y, LIN C, WANG Y, et al. Neural video compression with in-loop contextual filtering and out-of-loop reconstruction enhancement[C]//Proceedings of the 33rd ACM International Conference on Multimedia. [S.l.]: ACM, 2025: 12016-12024.

[75] WANG Y, HUANG Q, TANG B, et al. Multiscale motion-aware and spatial-temporal-channel contextual coding network for learned video compression[J]. Knowledge-Based Systems, 2025, 316: 113401.

[76] ZHANG C, SUN H, KATTO J. FLAVC: Learned video compression with feature level attention[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, TN, USA: IEEE, 2025: 28019-28028.

[77] BIAN Y, TANG C, LI L, et al. Augmented deep contexts for spatially embedded video coding[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, TN, USA: IEEE, 2025: 2094-2104.

[78] WEI X, LIN J, XU J, et al. RDVC: Efficient deep video compression with regulable rate and complexity optimization[J].

IEEE Transactions on Multimedia, 2025, 27: 5480-5491.

[79] MENTZER F, AGUSTSSON E, BALLÉ J, et al. Neural video compression using GANs for detail synthesis and propagation [C]//Proceedings of European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 562-578.

[80] YANG R, TIMOFTE R, VAN GOOL L. Perceptual learned video compression with recurrent conditional GAN[C]// Proceedings of the 31st International Joint Conference on Artificial Intelligence. Vienna: International Joint Conference on Artificial Intelligence, 2022: 1537-1544.

[81] DU P, LIU Y, LING N. CGVC-T: Contextual generative video compression with transformers[J]. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2024, 14(2): 209-223.

[82] LI B, LIU Y, NIU X, et al. Extreme video compression with pre-trained diffusion models[EB/OL]. (2024-02-14). https://arxiv.org/abs/2402.08934.

[83] MA W, CHEN Z. Diffusion-based perceptual neural video compression with temporal diffusion information reuse[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2025, 21(12): 1-22.

[84] MALL S, HENRIQUES J F. CRAM: Large scale video continual learning with bootstrapped compression[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2025: 15045-15055.

[85] LIU M, XU C, GU Y, et al. I^2VC: A unified framework for intra- & inter-frame video compression[EB/OL]. (2024-06-01). https://arxiv.org/abs/2405.14336.

[86] CHEN H, HE B, WANG H, et al. NeRV: Neural representations for videos[J]. Advances in Neural Information Processing Systems, 2021, 34: 21557-21568.

[87] LI Z, WANG M, PI H, et al. E-NeRV: Expedite neural video representation with disentangled spatial-temporal context[C]// Proceedings of European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 267-284.

[88] BAI Y, DONG C, WANG C, et al. PS-NeRV: Patch-wise stylized neural representations for videos[C]//Proceedings of 2023 IEEE International Conference on Image Processing (ICIP). [S.l.]: IEEE, 2023: 41-45.

[89] CHEN H, GWILLIAM M, LIM S N, et al. HNeRV: A hybrid neural representation for videos[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2023: 10270-10279.

[90] LEE J C, RHO D, KO J H, et al. FFNeRV: Flow-guided frame-wise neural representations for videos[C]//Proceedings of the 31st ACM International Conference on Multimedia. [S.l.]: ACM, 2023: 7859-7870.

[91] WU C, QUAN G, HE G, et al. QS-NeRV: Real-time quality-scalable decoding with neural representation for videos[C]// Proceedings of the 32nd ACM International Conference on Multimedia. [S.l.]: ACM, 2024: 2584-2592.

[92] KWAN H M, GAO G, ZHANG F, et al. HiNeRV: Video compression with hierarchical encoding-based neural representation [J]. Advances in Neural Information Processing Systems, 2023, 36: 72692-72704.

[93] LING Q, CHENG Z, FENG D, et al. A multi-grid implicit neural representation for multi-view videos[EB/OL]. (2025-09-20). https://arxiv.org/abs/2509.16706.

[94] GAO G, TENG S, PENG T, et al. GIViC: Generative implicit video compression[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2025: 17356-17367.

[95] JIA C, YE F, DONG F, et al. MPAI-EEV: Standardization efforts of artificial intelligence based end-to-end video coding[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 34(5): 3096-3110.

[96] DI TORINO I T, ARTUSI A, ANGELINI M, et al. Towards an AI-enhanced video coding standard[C]//Proceedings of International Broadcasting Convention. Amsterdam, Netherlands: IBC, 2022.

[97] GOMES J S, GRELLERT M, RAMOS F L L, et al. End-to-end neural video compression: A review[J]. IEEE Open Journal of Circuits and Systems, 2025, 6: 120-134.

[98] GUO H, ZHOU Y, GUO H, et al. A survey on recent advances in video coding technologies and future research directions[J]. IEEE Transactions on Broadcasting, 2025, 71(2): 666-671.

[99] 田港一, 纪雯. 机器视觉编码技术研究及其进展[J]. 计算机学报, 2025, 48(11): 2631-2665.
TIAN Gangyi, JI Wen. Research and advances in machine vision coding technology[J]. Chinese Journal of Computers, 2025, 48(11): 2631-2665.

[100] GAO W, LIU S, XU X, et al. Recent standard development activities on video coding for machines[EB/OL]. (2021-05-26). https://arxiv.org/abs/2015.12653.

作者简介:

何小海(1964-),男,教授,博士生导师,研究方向:图像/视频通信、机器视觉与智能系统,E-mail:hxh@scu.edu.cn。

李鑫磊(2002-),男,硕士研究生,研究方向:视频编码。

魏海涛(1997-),男,博士研究生,研究方向:图像/视频编码、机器视觉。

毕晓东(1996-),男,博士研究生,研究方向:图像/视频质量评价、视觉安全评价、视频编码。

聂尧佳(2000-),女,硕士研究生,研究方向:视频编码、机器视觉。

熊志娜(2001-),女,硕士研究生,研究方向:视频编码、VVC优化。

张皓彦(2002-),男,硕士研究生,研究方向:视频编码、码率控制。

熊淑华(1969-),**通信作者**,女,副教授,硕士生导师,研究方向:图像/视频编码、图像处理,E-mail:xiongsh@scu.edu.cn。

(编辑:王静)

Deep Learning-Driven Video Coding: Methods, Progress, and Perspectives

HE Xiaohai[1], LI Xinlei[1], WEI Haitao[1], BI Xiaodong[2], NIE Yaojia[1], XIONG Zhina[1], ZHANG Haoyan[1], XIONG Shuhua[1*]

(1. College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China;2. School of Cyber Science and Engineering, Sichuan University, Chengdu 610065, China)

Abstract: With the explosive growth of video data, limited network bandwidth and high computational demands pose significant challenges for video transmission and storage. In this context, the continuous development of efficient video coding methods is of critical theoretical significance and practical value, as it ensures the delivery of high-quality video services under resource-constrained conditions. However, traditional hybrid video coding frameworks have gradually reached performance bottlenecks, making further improvements in coding efficiency increasingly difficult. In recent years, deep learning, with its powerful nonlinear fitting and representation capabilities, has provided new opportunities for optimizing video coding. This paper presents a systematic and detailed analysis of deep learning-driven video coding technologies. First, we briefly introduce video coding techniques under conventional coding frameworks and further explore the optimization of key modules, such as intra- and inter-frame prediction, through deep learning. Then, we focus on the development and key technical routes of end-to-end video coding frameworks based on deep learning, providing a comparative analysis of their performance. Finally, we highlight significant research achievements of deep learning in the field of video coding, examine the challenges and limitations of existing techniques, and offer an outlook on future trends in video coding technologies.

Highlights:

1. This paper presents a systematic review of deep learning-driven video coding technologies, tracing the evolution of video coding from conventional frameworks to deep learning-based paradigms. It offers an in-depth analysis of two primary technical streams: Hybrid coding frameworks that integrate deep learning into traditional architectures, and fully end-to-end neural video codecs.

2. Building on an analysis of current trends and future demands in deep learning-based video coding, this paper identifies the key challenges that hinder the practical deployment of existing technologies and highlights several promising directions for future research.

Key words: video compression; deep learning; neural networks; learning-based video coding; end-to-end video coding

---

Foundation items: National Natural Science Foundation of China (Nos.62271336, 62211530110); The Key Research and Development Program of Sichuan Province (No.2024YFHZ0289); TCL Science and Technology Innovation Fund (No.0020506107005); The Key Research and Development Support Program of Chengdu (No.2024-YF06-00079-HZ).

Received: 2026-01-09; **Revised**: 2026-02-26

***Corresponding author**, **E-mail**: xiongsh@scu.edu.cn.