ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 2,Mar. 2025,pp. 303－319
DOI:10. 16337/j. 1004-9037. 2025. 02. 003

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

端到端智能视频压缩技术及其在无人机中的应用

叶　枫，董凡可，贾川民

（北京大学王选计算机研究所，北京 100871）

摘　要：多媒体视觉表示与传输领域正在面临深刻变革，端到端优化的智能视频编解码技术是激发这一变革的驱动力。以无人机（Unmanned aerial vehicle，UAV）视频为代表的新兴视频内容压缩编码技术进一步促进了核心技术发展和应用场景创新。聚焦于端到端智能视频编解码技术及其在无人机视频编码的初探，提出了一种基于分层双向参考结构的视频编码方法，解决模型在运动表示效率和预测编码精度方面的不足。有针对性地设计提出了参数共享的运动编解码器、双向缩放运动表示方法以及可信运动建模技术，显著提升无人机视频压缩的率失真压缩性能，优于传统视频编码标准H.266/VVC。为智能视频编码关键技术发展和应用提供了新思路，未来有望在无人机视觉感知等相关领域发挥重要作用。

关键词：端到端视频编码；编码标准；分层双向预测；无人机视频

中图分类号：TP37　　**文献标志码：**A

End-to-End Video Compression Technology and Its Application in Unmanned Aerial Vehicles

YE Feng，DONG Fanke，JIA Chuanmin

（Wangxuan Institute of Computer Technology，Peking University，Beijing 100871，China）

Abstract：The field of multimedia visual representation and transmission is undergoing profound transformation，with end-to-end optimized intelligent video coding technologies serving as the driving force. The compression of emerging video content represented by unmanned aerial vehicle（UAV）videos has further stimulated the development of core technologies and innovation in application scenarios. Focusing on end-to-end video coding technology and its initial exploration in UAV video coding，this study proposes a hierarchical bi-directional reference structure-based video coding method that addresses the shortcomings of existing models in motion representation efficiency and predictive coding accuracy. The targeted design introduces a parameter-shared motion codec，a bi-directional scaled motion representation method，and credible motion modeling technology，significantly improving the rate-distortion performance of UAV video compression and outperforming traditional video coding standards such as H.266/VVC. This work provides novel insights for the advancement of key intelligent video coding technologies and their practical applications，demonstrating promising potential for future deployment in UAV visual perception and related domains.

Key words: end-to-end video coding; coding standard; hierarchical bi-directional prediction; drone video

---

基金项目：北京市自然科学基金（4252003）；国家自然科学基金（62371008）。
收稿日期：2025-02-10；**修订日期：**2025-03-05

引　言

近年来，数字视频技术发展迅速，视频已经成为了人们记录和分享生活的主要途径之一，在教育、医疗、商业等多个领域都得到了广泛的应用，深刻影响着人们的学习、工作和娱乐方式。同时，视频会议、直播、远程桌面控制等新兴应用场景不断涌现，使得视频数据成为了信息传播的关键载体。这一趋势不仅促使着视频内容的产生和传播，还推动着视频编码技术的持续发展和革新，以应对海量视频数据存储和传播带来的挑战。

视频编码作为数字视频技术的基石，对于降低存储成本、提高传输效率起着至关重要的作用。一方面，深入研究视频编码的原理与算法，能够推动理论创新和方法突破，为学术研究提供新的可行方向。另一方面，高效的视频编码技术能够有效减少存储与带宽消耗，提升视频服务质量和用户体验，从而促进数字视频行业的可持续发展。同时，随着视频内容的指数级增长和新兴应用场景的不断拓展，市场对更先进、更智能的视频编码技术的需求日益增长，为相关企业和研究机构提供了广阔的发展机遇。因此，研究高效、智能的视频编码技术，对学术界和工业界均具有重要意义。

近年来，神经网络技术在计算机视觉领域取得了突破性进展，并在时序信息处理、图像特征增强等任务中展现出卓越的性能[1]。基于深度学习的视频编码方法已成为新的研究热点。早期工作尝试在传统编码模型中引入神经网络模块，来优化预测、环路滤波等关键组件，形成了一系列混合编码框架及标准，提高了压缩效率并降低了编码时间。然而，由于该混合结构仍然受到传统编码架构的限制，只能对模块进行单独改进，难以实现整体优化。因此，全神经网络的视频编码方法逐渐成为新的发展趋势。全神经网络编码通过端到端的训练和优化，能够提高编码工具的灵活性，通过统一的率失真函数来实现更高的压缩效率。此外，通过调整训练数据集，神经网络编码方法可以实现针对特定数据类型的自适应优化，无需繁琐的人工调参，使其能够适应更加复杂多变的应用场景。

在众多新兴应用场景中，无人机视频数据因其复杂性和多变性，成为智能编码技术的重要研究方向。无人机拍摄产生的大量图像、视频及传感器数据具有高维度、非结构化和动态变化的特点。同时，这些数据往往受到光照变化、天气干扰、运动模糊等因素的影响，传统的数据处理方法难以满足复杂环境下的实时性与准确性要求。根据商用无人机市场的预测，到2028年，全球商用无人机市场的规模将达到5 014亿元[2]。无人机搭载的摄像头在抢险救灾、科学探索及农业监测等领域展现出重要应用价值[3]。因此，提高无人机视频的压缩效率与编码速度已成为迫切需求，也使得视频编码技术在无人机领域的应用变得至关重要。为此，本文提出了一种新的端到端视频编码算法，该算法基于分层双向预测结构，通过参数共享的双运动编解码器对双向运动分别进行编码，并引入缩放运动先验和可信运动建模技术，提高预测特征的准确性，从而优化复杂数据场景下的编码效率。此外，合理的层间码率分配策略和高效的多阶段模型训练策略也进一步提升了模型的整体压缩性能。

1　相关工作

1.1　端到端视频编码

端到端视频编码作为一种新兴的编码技术，利用了神经网络强大的非线性建模能力，能够进一步提升运动估计、运动补偿和残差压缩等模块的性能，克服了传统基于块的混合编码框架的局限性。如图1所示，端到端视频编码经历了从像素域到特征域的转变和从残差编码到特征编码的演化，并伴随着多参考帧技术和双向预测方法两大核心技术的突破。

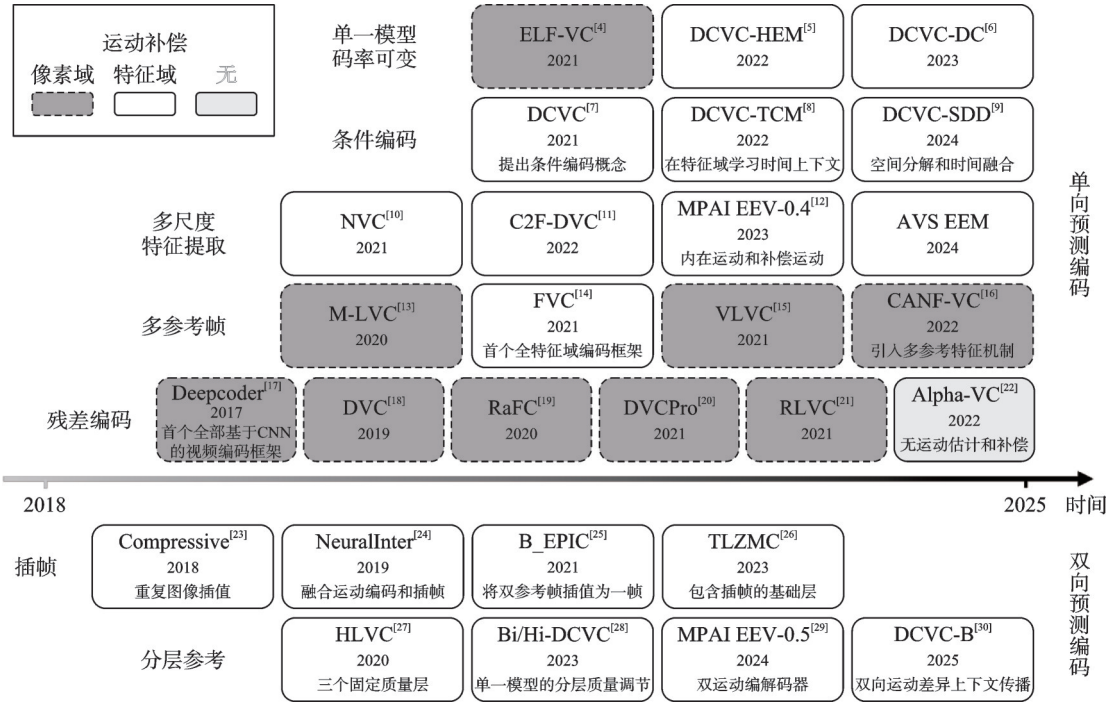


图1　端到端视频编码方法的技术路线图

Fig.1　Technical roadmap of end-to-end video coding methods

(1)从像素域到特征域的关键跨越

早期研究[17-18]通过像素级光流实现运动估计。这类方法依赖于精细的结构设计,面临复杂运动建模不准确与噪声敏感的固有局限,同时缺乏对前景/背景运动差异的语义理解。这一瓶颈促使研究者转向特征域操作:Feng等[31]首次在特征空间计算残差,减少运动误差传播带来的影响;Hu等[14]构建全特征域编码框架,在特征空间中完成运动估计、运动补偿和残差压缩操作;Sheng等[8]通过传播特征建模时序上下文,显著提升了语义运动理解能力。特征域操作为处理无人机视频中的动态背景分离与小目标运动捕捉奠定了技术基础。

(2)条件编码的提出与发展

传统残差编码受限于香农熵理论边界[7],而条件编码将条件定义为可学习的特征域上下文,能够利用丰富的信息辅助编码、解码和熵建模,实现更高效的时空信息利用。最初,条件编码被应用于特定模块,如熵编码和前景内容编码[32]。随后,Li等[7]通过在特征域中进行运动估计和运动补偿来生成上下文特征,显著增强了条件信息的容量和能力。条件编码的潜力在CANF-VC[16]中得到了进一步体现,将其扩展到运动编码,形成了一个完全基于条件编码的视频压缩框架,实现了良好的编码性能,凸显了其相对于残差编码方法的显著优势。

(3)多参考帧/特征的引入

与单个参考帧相比,额外的参考帧可以提供更多的运动和纹理等信息,提高运动估计和运动补偿的精度,增强预测性能,显著减少需要编码的残差或条件信息,从而降低码率,提升编码性能。为了充分利用多个历史帧的信息,Hu等[14]引入了基于非局部注意力机制的多特征融合模块,以辅助当前帧的重建。Ho等[16]引入了多参考特征机制,通过一个流外推网络,从多个参考帧和特征中生成一张流图,进而获得更准确的预测信息作为条件辅助后续编码。多参考帧使得编码器能够更好地适应多样化的

场景和内容,对于无人机视频具有的复杂运动、遮挡等特性,多参考帧能够带来显著的性能提升。

(4)双向预测编码的探索

与单向预测编码算法相比,基于双向预测的神经网络编码方法受到的关注较少,但仍然是视频编码领域中一个重要的发展方向,具有独特的发展轨迹。通过帧插值[23-26]和分层编码结构[27-28]技术,双向预测编码算法实现了高效的编码性能和灵活的编码策略。

2024年,分层双向预测编码算法取得了显著进展。Ye等[29]提出了一种全新的双向编码框架,采用两个参数共享的运动编解码器,有效利用了时间上下文中的缩放运动信息作为先验。该模型结合了可信的运动建模和高效的码率分配策略,以优化整体压缩性能。此外,Sheng等[30]开发的DCVC-B模型采用了双向运动差异上下文传播方法,有效提升了压缩性能。这些进展使得神经网络编码模型在相同配置下,在RGB和YUV420色彩空间中的部分数据集上超越了H.266/VVC参考软件的性能。

然而,如表1所示,早期方法大多基于插帧模型,没有进行精细的特征提取,难以实现良好的编码性能。近期方法虽然进行了显式的运动编码,但在动态层级分配与运动复杂度感知方面仍存在不足,难以适配无人机视频的非平稳运动特性。这为本研究提出的分层双向编码框架提供了创新空间。

表1 双向预测编码模型总结

Table 1 Summary of bi-directional prediction coding models

方法	运动编码器数量	插帧	分层参考结构	特征提取
Compressive[23]	0	√	手工设计	×
NeuralInter[24]	1	√	自适应	×
HLVC[27]	1	×	手工设计	×
B_EPIC[25]	1	√	自适应	×
TLZMC[26]	1	√	自适应	×
Bi/Hi-DCVC[28]	1	×	自适应	√
DVC-SHBMM[29]	2	×	自适应	√
DCVC-B[30]	1	×	自适应	√

1.2 无人机视频编码

在过去的10年中,无人机因其灵活性和强大的空间感知能力,受到越来越多的关注。随着传感器技术的进步,搭载在无人机上的摄像头生成的视频数据量和比特率急剧增加,对存储以及空对地数据传输提出了更高的要求。目前的传统视频编码方法主要针对自然场景设计,未能充分考虑无人机视频中独特的纹理和视角特征。

无人机视频具有独特的特点,使其与自然视频有所不同,也为视频编码带来了新的挑战。无人机视频通常是由运动中的无人机摄像头从不同视角和高度拍摄,视频中往往存在大范围的运动、视角畸变(如鸟瞰或鱼眼视角)和复杂的背景变化。此外,无人机视频的分辨率一般较高(如2 720像素×1 520像素)且帧率通常为24帧/s~30帧/s,这对编码算法的计算复杂度和编解码时间提出了更高要求。无人机视频涵盖的场景种类丰富,包括道路交通、城市与乡村区域、室内外环境等,而视频中的目标对象(如行人、车辆)密度各异,进一步增加了编码的复杂性。由于无人机的快速移动,视频中常出现运动模糊和目标遮挡现象,这对运动估计与补偿提出了更高的挑战。

目前,针对无人机视频的编码方法通常基于传统视频编码标准实现,如H.264、H.265及其扩展版本,通过构建辅助模块或特定优化算法来适应无人机视频的特征[33-34]。然而,这些方法在处理无人机视频时存在明显的局限性。传统的基于块的运动补偿预测方法难以准确捕捉无人机视频中的复杂运动,

尤其是在处理大范围运动或视角畸变时表现不佳。同时,由于无人机视频的分辨率较高,现有方法在压缩过程中会消耗大量比特,导致存储和传输成本上升。更重要的是,现有方法未能充分考虑无人机视频的特有特性,使得压缩效率大打折扣。

相比于传统的基于块的运动补偿方法,端到端视频编码模型在无人机视频编码中展现了显著优势。基于神经网络的编码方法能够通过细粒度的像素级运动模型灵活地描述无人机视频中的复杂运动,尤其在应对大范围运动或视角畸变时表现尤为出色。通过端到端的联合优化,端到端视频编码模型能够学习到无人机视频的内在规律,减少对手工特征设计的依赖,从而提升压缩效率。此外,端到端视频编码模型能够从广泛的训练集中学习到无人机视频的内容特性,从而增强对复杂场景的鲁棒性。

2　端到端智能视频编码方法

面对无人机视频高动态、多尺度运动特性带来的压缩挑战,传统视频编码框架难以实现良好的压缩效果。为此,本节从无人机视频的复杂特性出发,创新性地提出了一个基于分层双向参考结构的端到端视频编码方法,以弥补现有编码方法对于无人机视频压缩效果的不足。首先,在2.1节揭示现有端到端视频编码方法在分层双向预测领域研究的不足,确立本方法的必要性;进而在2.2节提出算法的核心框架,并在2.3~2.6节对运动编码与信息融合、缩放运动先验、可信运动建模和率失真优化这4项模块和技术进行了详细介绍,阐述其对于处理无人机视频独特特性的作用;最后在2.7节描述了模型优化及标准化探索工作,为未来的进一步研究指明方向。

2.1　研究动机

传统视频编码标准(如H.264/AVC、H.265/HEVC 和 H.266/VVC)在随机访问配置下通常采用分层双向预测结构,以提升视频压缩效率。这种结构通过对视频帧进行层级划分,使得低层帧具有更高质量,为其他帧提供更可靠的参考信息,从而增强整体编码性能。

然而,端到端优化的神经网络编码方法在该框架下的研究仍较为有限。现有的端到端视频编码方案主要关注单向预测结构,而对分层双向预测的优化策略、参考信息的高效利用以及运动估计的可靠性缺乏系统性的研究。

为弥补这一不足,本研究提出了一种基于分层参考结构的双向预测编码模型,并结合参数共享的运动编解码器、双向缩放运动先验以及可信运动建模等关键技术,优化前后参考信息的融合方式,提高整体压缩效率。提出的模型还支持通过超参数对层间码率分配比例进行调节,使得模型能够更好地应对不同的数据特点和视频内容。此外,为确保模型的稳定性和适应性,还设计了一整套有效的训练策略,以提升在不同视频场景下的编码性能。

本研究提出的方法已被基于人工智能的视频、音频和数据编码-端到端优化神经视频编码小组(Moving picture, audio, and data coding by artificial intelligence-end-to-end optimized neural video coding, MPAI-EEV)项目采纳,作为其最新的参考模型EEV-0.5,进一步推动了端到端视频编码标准的发展。

2.2　算法框架

算法的核心是一个分层双向预测结构,对前后参考帧进行独立的运动编码,结合双向缩放运动先验、高效的可信运动建模与信息融合策略,能够有效捕捉帧间的时间相关性,并在管理参考帧时考虑质量变化,还通过灵活的码率分配策略提高压缩性能。如图2所示,算法框架分为3个主要步骤:先验信息生成、预测特征融合和残差编码。首先,通过光流网络估计前向和后向运动向量,并分别进行运动编码、解码和补偿,得到双向预测特征;接着,采用高效的信息融合策略,获得精确的先验特征;最后,通过

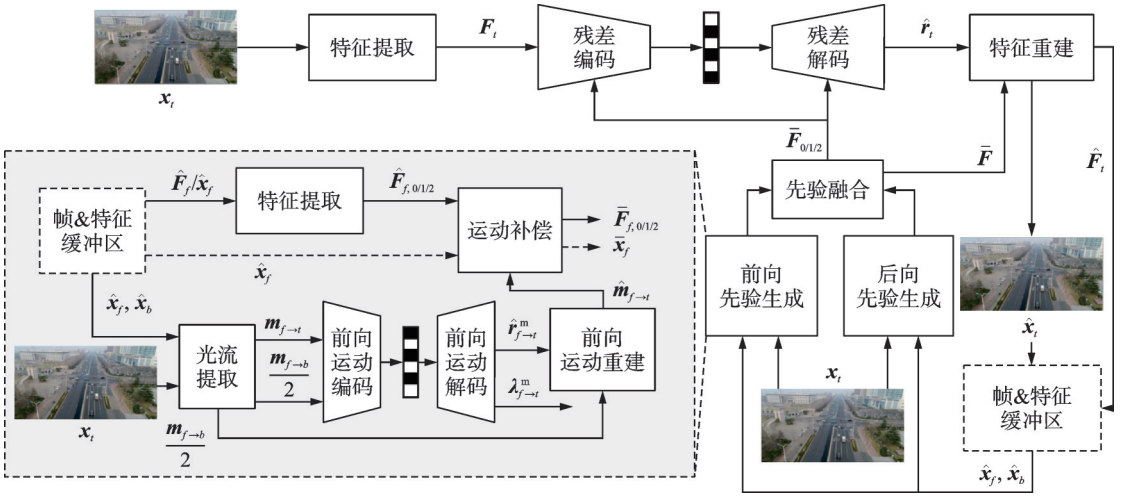


图2 提出的分层双向预测编码算法的框架图

Fig.2 Framework of the proposed hierarchical bi-directional prediction coding algorithm

残差编码器、解码器和特征重建模块生成重建帧。该模型能够在保持高预测精度的同时，降低计算和存储开销，从而推动视频编码技术在实际应用中的发展。

2.3 运动编码与信息融合

对于无人机视频序列来说，往往存在着尺度缩放、镜头旋转等复杂的运动，难以进行精确的建模。同时，双向的运动往往并不具有对称性。以往的方法都使用一个运动编解码器，同时对双向运动进行编码，这一设计不仅不能减小运动编码阶段的码率，还会影响运动预测的准确性。因此，为了更准确地表示运动信息，本文提出利用2个参数共享的运动编解码器，对前、后向运动分别进行编码，结构如图3所示。图中的Spynet为光流网络，Q、AE和AD分别表示量化(Quantization)、算术编码(Arithmetic encoding)和算数解码(Arithmetic decoding)。RB表示深度块(Residual block)，Conv表示卷积层，通过控制这两种结构中的参数，可以实现尺度的上采样(↑2)和下采样(↓2)。

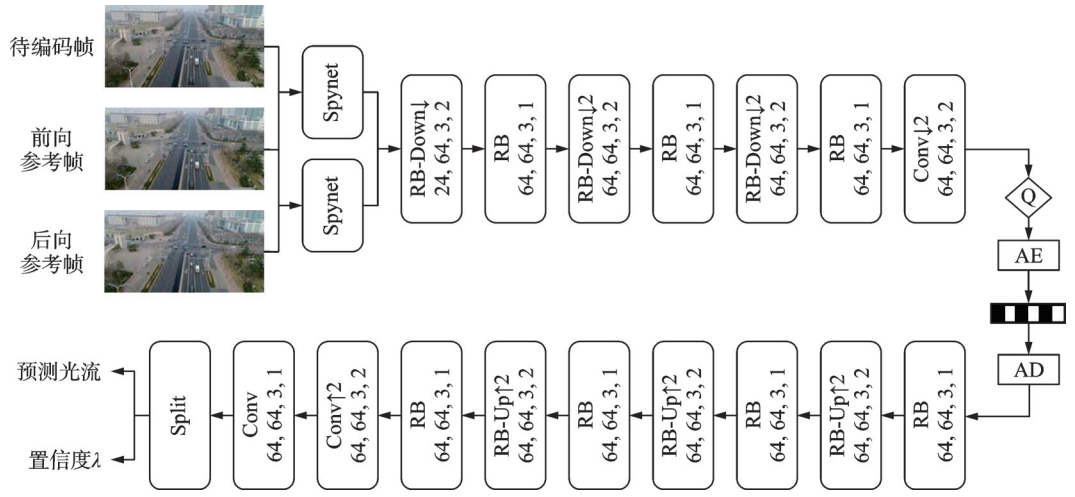


图3 运动编解码器的结构

Fig.3 Architecture of the motion codec

这种设计不仅能够直接迁移单向编码方案中的高效运动编码模块,还能通过额外的运动编解码器获得更精确的运动预测信息,从而更好地理解无人机视频的复杂的时间域特性。

在信息融合方面,算法从参考帧或参考特征和重建的运动向量中提取3层特征,特征提取器的结构如图4所示。然后,在每个尺度上分别进行对齐和运动补偿,获得单侧的前向预测特征 $\bar{F}_{f,0}$、$\bar{F}_{f,1}$、$\bar{F}_{f,2}$ 和后向预测特征 $\bar{F}_{b,0}$、$\bar{F}_{b,1}$、$\bar{F}_{b,2}$。最后,将前向和后向的输出进行拼接融合,生成最终的预测特征 $\bar{F}_0$,$\bar{F}_1$,$\bar{F}_2$,作为后续残差编码的先验信息。这种框架设计不仅提供了准确可靠的先验信息,还使得残差编码过程与低延迟框架的设计保持一致,简化了现有端到端模型的迁移。

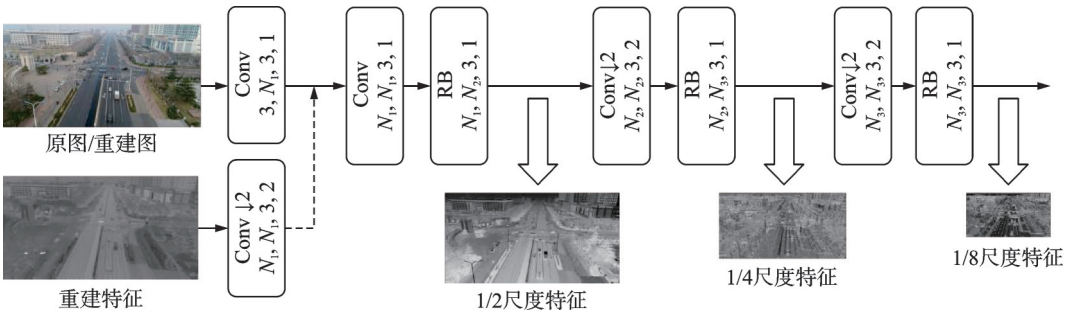


图4　特征提取器的结构

Fig.4　Architecture of the feature extractor

2.4　缩放运动先验

为了减少运动编解码器带来的额外比特开销,算法利用前向和后向参考帧之间的光流信息,进一步优化双向预测过程。根据光流方法的基本假设(亮度恒定、时间连续或运动幅度小),可以推测待编码帧与参考帧之间的光流应与前向和后向参考帧之间的光流具有紧密的时间相关性。尤其是对于运动平缓的区域(如背景),待编码光流与参考光流几乎相同,这为运动信息的压缩提供了天然的优势。

具体而言,以前向运动编码过程为例,算法计算待编码帧与前向参考帧之间的光流 $\boldsymbol{m}_{f\to t}$,以及后向参考帧与前向参考帧之间的光流 $\boldsymbol{m}_{f\to b}$。将 $\boldsymbol{m}_{f\to b}$ 的一半作为先验,与 $\boldsymbol{m}_{f\to t}$ 拼接后输入运动编码器。在解码时,运动解码器的输出 $\hat{\boldsymbol{r}}_{f\to t}$ 与 $\boldsymbol{m}_{f\to b}/2$ 共同输入运动信息重建模块,生成最终的重建运动 $\hat{\boldsymbol{m}}_{f\to t}$。

这种基于缩放运动先验的设计,能够有效减少运动编码的码率,尤其适用于线性和幅度较小的运动场景。以往方法选择直接编码 $\boldsymbol{m}_{f\to t}$ 和 $\boldsymbol{m}_{f\to b}/2$ 的差值[26,29],但是对于无人机视频序列中的非线性和剧烈运动,这一设计不仅起不到降低码率的作用,还无法充分利用参考光流中的信息来应对遮挡、运动模糊等问题。本文提出的将参考光流作为先验的设计思路能够更加有选择性地利用时间上下文中的缩放运动信息,使得模型能够更好地捕捉复杂运动模式,显著提升运动估计的精度和鲁棒性,减少运动补偿误差,从而降低后续残差信息的编码成本。

2.5　可信运动建模

在分层参考结构中,两个参考帧通常位于不同的质量层,其重建质量可能存在较大差异。此外,无人机视频中的相机旋转、剧烈运动和遮挡等情况可能导致某一侧的参考帧无法提供有效的参考信息。为此,本文提出了可信运动建模的概念,用于动态评估参考信息的有效性,从而优化预测过程。

以前向运动编码过程为例,运动解码器的输出被扩展为两部分:预测运动信息 $\hat{\boldsymbol{r}}_{f\to t}$ 和置信度 $\boldsymbol{\lambda}_{f\to t}$。通过Sigmoid函数将 $\boldsymbol{\lambda}_{f\to t}$ 的值归一化到(0,1)范围内。最终的预测帧 $\bar{\boldsymbol{x}}_t$ 由前向和后向预测帧的加权生成,即 $\bar{\boldsymbol{x}}_t=\boldsymbol{\lambda}_{f\to t}\cdot\bar{\boldsymbol{x}}_f+\boldsymbol{\lambda}_{b\to t}\cdot\bar{\boldsymbol{x}}_b$。类似的,在特征融合阶段,$\boldsymbol{\lambda}_{f\to t}$ 通过卷积神经网络进行扩展并下采样,得到多尺度的置信度 $\boldsymbol{\lambda}_{f\to t,0}$、$\boldsymbol{\lambda}_{f\to t,1}$ 和 $\boldsymbol{\lambda}_{f\to t,2}$。这些置信度分别与对应尺度的前向预测特征相乘,后向预测

特征也进行相同操作，最终生成多尺度预测特征 $\bar{F}_0$、$\bar{F}_1$ 和 $\bar{F}_2$，用于后续的残差编码。残差编解码器的结构如图5所示。

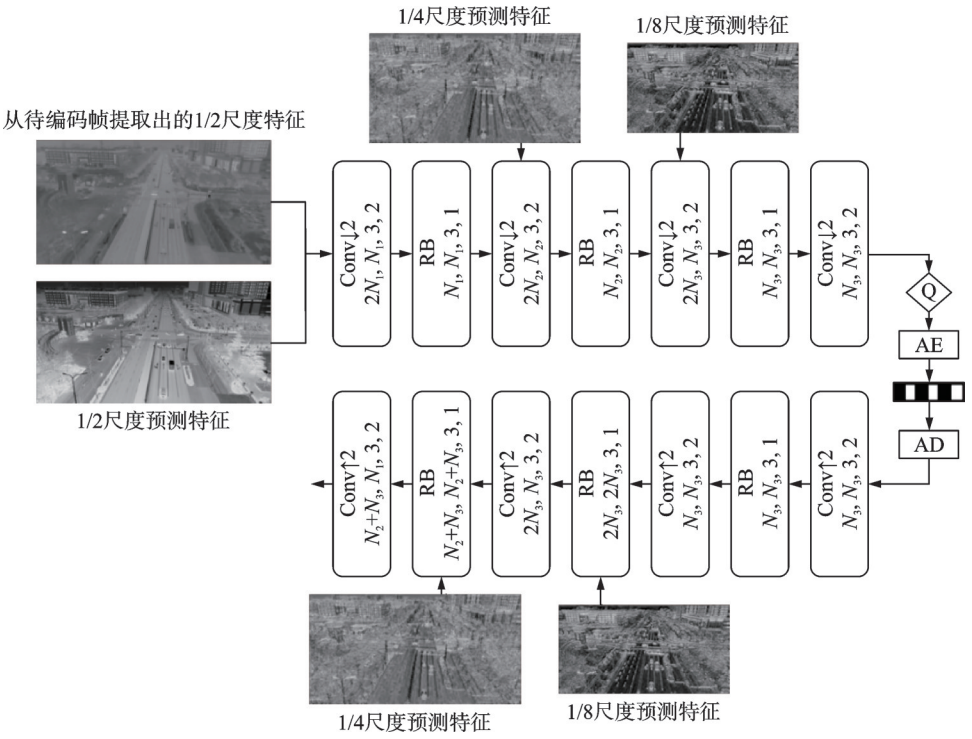


图5 残差编解码器的结构

Fig.5 Architecture of the residual codec

这种基于置信度的加权融合机制能够有效应对参考帧质量不一致或参考信息不可靠的情况，从而提升运动补偿的精度和鲁棒性。通过动态评估参考信息的有效性，模型能够更好地适应复杂场景（如剧烈运动或遮挡），并生成更高质量的预测帧和特征，最终降低残差编码的复杂度并提升整体压缩效率。

2.6 率失真优化技术

在分层结构中，合理的码率分配策略对整体压缩性能至关重要。为了在同一模型中实现不同层次B帧的压缩质量，算法通过优化编解码器结构和损失函数来实现动态码率分配。具体而言，算法引入了一种灵活的码率控制机制，能够根据视频内容的特点和压缩需求，自适应地调整各层次B帧的码率分配。

首先，算法添加了一个全连接函数标量，其输入为[0, 1.4]范围内的浮点数 λ_{scalar}，输出为一个多维向量。该向量根据编解码器和熵模型的维度进行分组，并与神经网络特定层的输出逐元素相乘，从而实现对编码过程中各层次特征的动态调整。这种设计使得模型能够根据 λ_{scalar} 的值灵活控制不同层次B帧的压缩质量，例如在高质量层分配更多码率以保留细节信息，而在低质量层减少码率以提升压缩效率。

此外，λ_{scalar} 还与损失函数中的原始 λ 相乘，进一步优化码率分配策略。通过这种双重调节机制，模型能够在训练过程中动态平衡码率与失真之间的关系，从而在保证视觉质量的同时最大化压缩效率。

表2展示了不同码率点下λ_{scalar}的取值。

图6展示了提出的模型在同一序列的高码率点和低码率点上的帧质量和比特率对比。可以发现,相较于低码率点,虽然高码率点在进行层间码率调节时,已经给高质量层分配了更高比例的码率,但是从峰值信噪比(Peak signal-to-noise ratio, PSNR)指标来看,其质量分层效率仍然不如低码率点明显,这说明了为不同码率点分别设计λ_{scalar}取值的合理性。

表2　不同码率点下的λ_{scalar}取值

Table 2　Values of λ_{scalar} at different bit rate levels

码率点	λ_{scalar}			
	第1层	第2层	第3层	第4层
λ=2 048	0.80	0.40	0.25	0.1
λ=1 024	0.80	0.40	0.25	0.1
λ=512	0.75	0.35	0.25	0.1
λ=256	0.70	0.30	0.20	0.1

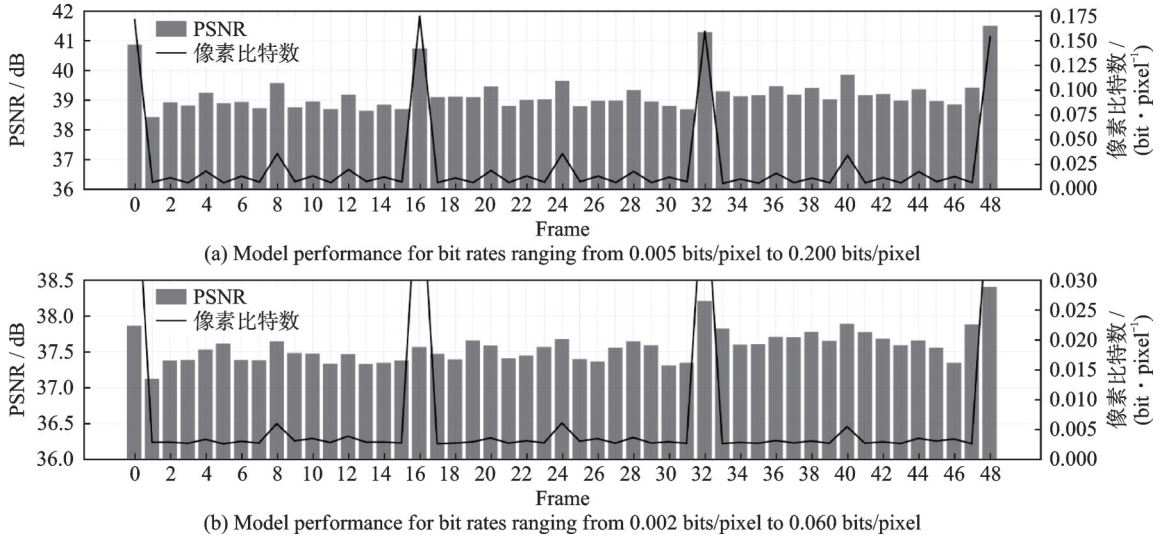


图6　提出的模型在电梯序列上的帧质量和比特率展示

Fig.6　Frame quality and bitrate performance of the proposed model on elevator sequence

特别的是,提出的基于超参数的层间码率调节方法,在模型训练完成之后,仅在测试阶段根据数据特点修改λ_{scalar}的值,仍然能够实现有效的调节。这一动态调节能力使得模型能够针对无人机视频的不同特性和场景进行较好的适应。例如,对于运动非常剧烈、视角切换较快的序列,本模型可以在支持可变帧内周期的同时,任意调节层间码率分配比例,有利于提高整体的编码性能。

2.7　模型优化及标准化探索

为了追求更高的性能,本研究还对编码方法进行了多方面的优化,特别是在通道数设计和I帧模型选择上进行了重要改进。在多层特征提取阶段,采用了渐进式通道数增加的策略,随着图像尺度的逐渐减小,显著增加了每一层的通道数,这不仅能够更有效地捕捉到细粒度的特征,也有助于提升模型对复杂场景的适应能力。在低尺度特征图中,通道数的增加使得模型能够更精确地建模复杂运动和高频纹理信息,从而显著提升编码效率和质量。其次,通过增加通道数,不仅能够提高模型的表达能力,还能够充分测试模型的潜力。此外,还引入了更先进的I帧模型用于学习型图像压缩的线性复杂性多参考熵建模(Linear complexity multi-reference entropy modeling for learned image compression, MLIC++)[35],尽管该模型的计算量较大,但其在图像重建质量和细节保留方面具有显著优势。

MPAI是一个致力于利用人工智能提升多媒体数据压缩和处理效率的国际标准化组织。MPAI-EEV作为该组织的一部分,专注于研究端到端神经网络视频编码方法,旨在突破传统视频编码标准的

局限,实现更高效的视频压缩。自 2021 年 12 月启动以来,MPAI-EEV 项目已发布了 5 个版本的验证模型(EEV-0.1 至 EEV-0.5),并在自然场景视频编码和无人机视频编码中都取得了显著的编码效率提升。本章 2.3~2.6 节中提出的关键技术已成功集成到 EEV-0.5 参考软件中,为其功能完善和性能提升做出了重要贡献。这些技术显著提升了编码效率与视觉质量,尤其是在处理复杂运动和高分辨率视频时表现突出。该系列技术演进既验证了端到端视频编码框架的实用性,也为未来更大规模模型的引入和码率分配策略的优化奠定了创新路径,持续推动视频压缩技术的进步。

本章提出的分层双向参考编码框架,为解决无人机视频的镜头畸变、尺度变化和运动模糊等压缩难题提供了可行思路,其标准化探索和技术提案采纳也验证了提出方法的可行性。

3 无人机视频编码测试结果

为了验证提出算法的先进性,本节引入一个专为无人机视频编码设计的基准测试集,对其进行详细介绍和分析,并对现有传统编码方法和端到端编码方法进行广泛测试,展示了详细的测试结果。此外,3.4 节还对当前端到端视频编码在无人机领域的需求和不足进行了详细分析,提供了可能的发展方向。

3.1 测试集介绍

本文实验所使用的测试数据集是一个专门为无人机视频编码任务构建的基准测试数据集[36]。该数据集旨在解决无人机视频处理中的独特挑战,并为端到端视频编码模型提供多样化的测试场景。该数据集视频序列特性如表 3 所示。

表 3 无人机视频编码基准的视频序列特征

Table 3 Video sequence characteristics of UAV video coding benchmark

来源	序列名称	空间分辨率(像素×像素)	总帧数	帧率/(帧·s^{-1})	比特深度	场景特征
A类 VisDrone-SOT TPAMI2021[2]	篮球场	960×528	100	24	8	室外
	草地	1 344×752	100	24	8	室外
	交叉路口	1 360×752	100	24	8	室外
	夜间购物广场	1 920×1 072	100	30	8	室外
	足球场	1 904×1 056	100	30	8	室外
B类 VisDrone-MOT TPAMI2021[2]	环形公路	1 360×752	100	24	8	室外
	立交桥	2 720×1 520	100	30	8	室外
	高速公路	1 344×752	100	24	8	室外
C类 Corridor IROS2018[37]	教室	640×352	100	24	8	室内
	电梯	640×352	100	24	8	室内
	大厅	640×352	100	24	8	室内
D类 UAVDT_S ECCV2018[38]	校园	1 024×528	100	24	8	室外
	海边道路	1 024×528	100	24	8	室外
	剧院	1 024×528	100	24	8	室外

(1)数据集的多样性与特点

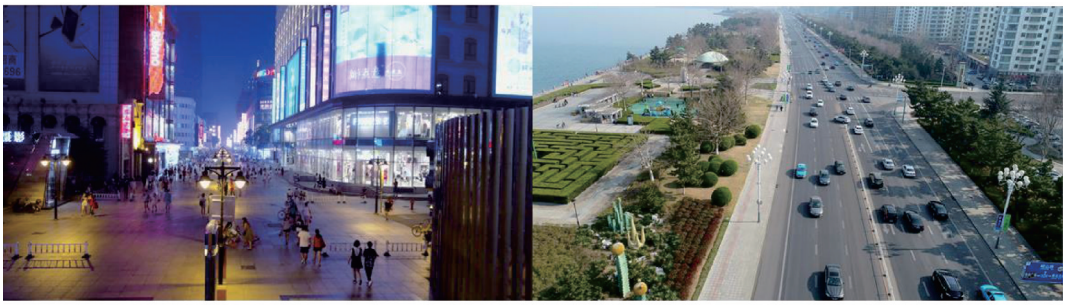
该数据集包含 14 个视频片段,分辨率范围为 640 像素×352 像素至 2 720 像素×1 520 像素,帧率为

24～30帧/s,每个视频片段包含100帧。这些片段从多个公开的无人机视频数据集中精选而来,涵盖了多样化的内容。具体特点包括:设备多样性,视频由不同型号的无人机摄像头拍摄,确保了设备类型的多样性;场景多样性,视频拍摄于多种地理位置和环境,包括室内(如教室)、室外(如篮球场、高速公路、校园)以及城市与乡村区域;目标对象多样性,视频中包含多种目标对象,如行人、车辆等,且目标密度从稀疏到拥挤不等;条件多样性,视频在不同光照条件、飞行高度和摄像头视角下采集,涵盖了无人机视频的典型挑战,如运动模糊、尺度变化和复杂背景。

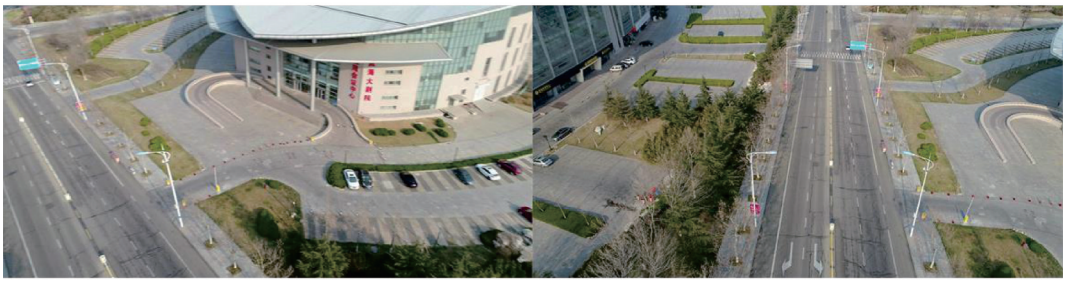
(2)无人机视频特性

图7展示了无人机视频数据集中有代表性的部分序列。

(a) Lens distortion



(b) Complex backgrounds and occlusion



(c) Widespread movement (camera rotation)

图7　无人机视频特性示例图

Fig.7　Examples of typical drone videos

运动模糊与尺度变化:由于无人机摄像头的快速运动,视频中常出现明显的运动模糊和目标尺度变化。运动模糊主要是由于无人机在飞行过程中快速移动或旋转,导致拍摄目标在帧间出现伪影,降低了图像的清晰度。同时,无人机在飞行高度和角度的变化会导致目标尺度发生显著变化,这种尺度变化使得目标特征的提取和匹配变得更加困难,进一步增加了视频处理的复杂度。

复杂背景与遮挡:无人机拍摄的视频通常具有复杂的背景,例如城市建筑、森林或水域等,这些背景中可能包含大量细节和纹理信息。此外,目标对象在视频中常受到遮挡或非刚性形变的干扰。例如,无人机拍摄的车辆或行人可能会被树木、建筑物或其他物体部分遮挡,或者由于目标的非刚性形变(如人体动作或旗帜飘动)导致目标外观发生变化。这些因素使得目标的跟踪和运动估计变得更加复杂,降低了传统视频处理算法的鲁棒性。

摄像头畸变:无人机摄像头通常采用鸟瞰或鱼眼视角,这种广角镜头在视角边界处会引入显著的畸变。这种畸变会导致视频中的直线在边缘区域变得弯曲,目标形状和比例失真,进一步增加了视频处理的难度。特别是在目标检测、跟踪和运动估计等任务中,畸变会严重影响算法的准确性,需要额外的校正和补偿步骤来消除畸变的影响。

帧间预测困难:由于无人机视频中视角和尺度的频繁变化,传统的基于块运动补偿的混合视频编码方法在处理无人机视频时效率较低。此外,目标尺度的变化也会使得块匹配的精度下降,从而降低了压缩效率并增加了码率开销。

(3)数据集的应用价值

该数据集不仅为无人机视频编码任务提供了高质量的测试基准,还为研究端到端视频编码模型在复杂场景下的性能提供了重要支持。通过在该数据集上的实验,可以全面评估模型在处理运动模糊、尺度变化、复杂背景和摄像头畸变等方面的能力。因此,本研究选择采用该数据集对模型的编码性能进行评估和对比,从而展现端到端视频编码技术在复杂场景下的潜能。

3.2 实验设置

对每个测试序列,取前96个重建帧进行质量对比,帧内周期(Intra-period)和图像组(Group of pictures,GOP)大小均设置为16。视频质量通过RGB色彩空间上的PSNR和多尺度结构相似度指数(Multi-scale structural similarity index measure,MS-SSIM)进行评估,压缩效率通过Bjøntegaard率失真增益(Bjøntegaard delta rate,BD-rate)进行衡量。

传统视频编码标准方面,选择H.265/HEVC(HM-16.20)和H.266/VVC(VTM-23.0)进行对比,分别测试其在低延时(Low delay P,LDP)配置和随机访问(Random access,RA)配置下的编码性能,以进行更加完善的对比。

端到端视频编码方法方面,选择EEV-0.1、EEV-0.4和EEV-0.5进行测试,以展示MPAI-EEV编码标准在性能优化方面的持续进步。同时,还提供了2.7节中提到的更高计算量的模型(记为EEV-0.5-L)的测试结果,展示提出的分层双向参考模型的发展潜力。

3.3 实验结果

表4和表5分别展示了PSNR和MS-SSIM指标下的测试结果,基准方法为HM-LDP。图8展示了PSNR指标下部分序列的性能对比图(码率-失真曲线),图9展示了PSNR指标下剧院、足球场和电梯序列的主观质量对比图。

对于PSNR和MS-SSIM测试指标,EEV-0.5的平均编码效率相较于EEV-0.1和EEV-0.4均有着明显的提升,特别是在PSNR指标上,EEV-0.5相较于上一版参考模型EEV-0.4,BD-Rate指标提升了超过34%。此外,不难发现EEV-0.5在基准方法为HM-LDP时,其性能增益十分稳定,避免了前几版参考软件在某些特定序列上性能显著下降的问题,进一步证明了其鲁棒性和通用性。

具体而言,在PSNR指标上,EEV-0.5的平均性能已经超越了VTM-RA。对于更大规模的模型EEV-0.5-L,在C类室内序列,尤其是教室序列上,取得了显著的性能提升。这主要得益于其更强的特征提取能力,使得EEV-0.5-L在处理镜头畸变严重的场景时表现尤为优异。

表4　RGB-PSNR指标下的测试结果(BD-rate)

Table 4　Test results for RGB-PSNR metrics(BD-rate)　%

来源	序列名称	HM-RA	VTM-LDP	VTM-RA	EEV-0.1	EEV-0.4	EEV-0.5	EEV-0.5-L
	篮球场	−32.02	−42.01	−65.64	44.00	−27.11	−73.39	−75.46
A类	草地	−31.17	−51.29	−73.93	−25.02	−51.61	−70.08	−71.94
VisDrone-SOT	交叉路口	−34.49	−55.83	−76.96	−12.41	−54.54	−77.74	−79.22
TPAMI2021[2]	夜间购物广场	−32.04	−47.63	−70.79	19.39	−42.19	−74.83	−76.25
	足球场	−35.56	−57.27	−77.30	16.39	−44.68	−74.25	−76.72
B类	环形公路	−33.99	−53.17	−74.06	−7.53	−36.34	−74.69	−75.73
VisDrone-MOT	立交桥	−33.53	−61.13	−79.39	71.23	−6.37	−68.84	−69.27
TPAMI2021[2]	高速公路	−31.20	−52.16	−73.84	8.26	−40.09	−74.34	−75.82
C类	教室	−29.74	−58.46	−70.97	63.97	−39.01	−68.85	−75.70
Corridor	电梯	−25.79	−50.36	−63.59	28.92	−41.27	−71.73	−75.40
IROS2018[37]	大厅	−33.07	−53.74	−71.36	7.47	−48.13	−77.82	−80.70
D类	校园	−34.07	−55.74	−77.60	2.48	−50.31	−79.90	−81.10
UAVDT_S	海边道路	−33.44	−54.01	−76.21	−3.72	−44.76	−77.36	−78.42
ECCV2018[38]	剧院	−37.77	−56.98	−79.12	33.57	−29.69	−72.57	−74.04
A类		−33.05	−50.80	−72.92	8.47	−44.02	−74.06	−75.92
B类		−32.91	−55.49	−75.76	23.99	−27.60	−72.63	−73.61
C类		−29.53	−54.19	−68.64	33.45	−42.80	−72.80	−77.27
D类		−35.09	−55.58	−77.64	10.78	−41.58	−76.61	−77.85
平均		−32.70	−53.56	−73.63	17.64	−39.72	−74.03	−76.13

表5　RGB-MS-SSIM指标下的测试结果(BD-rate)

Table 5　Test results for RGB-MS-SSIM metrics(BD-rate)　%

来源	序列名称	HM-RA	VTM-LDP	VTM-RA	EEV-0.1	EEV-0.4	EEV-0.5	EEV-0.5-L
	篮球场	−43.42	−44.40	−70.42	4.73	−49.23	−76.74	−81.95
A类	草地	−30.99	−61.42	−73.47	−54.31	−70.38	−69.18	−83.36
VisDrone-SOT	交叉路口	−44.01	−57.51	−78.27	−49.96	−71.72	−80.32	−87.13
TPAMI2021[2]	夜间购物广场	−50.67	−46.53	−77.01	−33.28	−61.60	−77.38	−83.80
	足球场	−47.97	−59.58	−79.10	−55.47	−66.36	−76.25	−85.30
B类	环形公路	−50.94	−57.98	−79.58	−52.36	−69.13	−82.68	−87.19
VisDrone-MOT	立交桥	−46.21	−44.59	−81.19	−18.44	−45.78	−71.09	−78.99
TPAMI2021[2]	高速公路	−40.51	−55.25	−76.52	−33.32	−59.94	−74.98	−81.16
C类	教室	−73.46	−85.60	−88.14	−32.49	−61.67	−65.97	−75.19
Corridor	电梯	−54.68	−79.79	−77.91	−57.37	−80.74	−80.08	−86.47
IROS2018[37]	大厅	−58.95	−77.71	−81.58	−51.84	−73.56	−77.37	−84.28
D类	校园	−43.82	−51.61	−78.54	−41.51	−64.56	−79.43	−84.08
UAVDT_S	海边道路	−45.32	−54.15	−78.91	−39.97	−64.09	−77.55	−81.40
ECCV2018[38]	剧院	−45.00	−59.72	−79.59	−20.14	−51.73	−69.40	−77.81
A类		−43.41	−53.85	−75.65	−37.66	−63.86	−75.97	−84.31
B类		−45.89	−52.60	−79.10	−34.70	−58.29	−76.25	−82.44
C类		−62.36	−81.03	−82.54	−47.23	−71.99	−76.67	−81.98
D类		−44.72	−55.16	−79.02	−33.87	−60.13	−75.46	−81.10
平均		−48.28	−59.70	−78.59	−38.26	−63.61	−75.60	−82.72

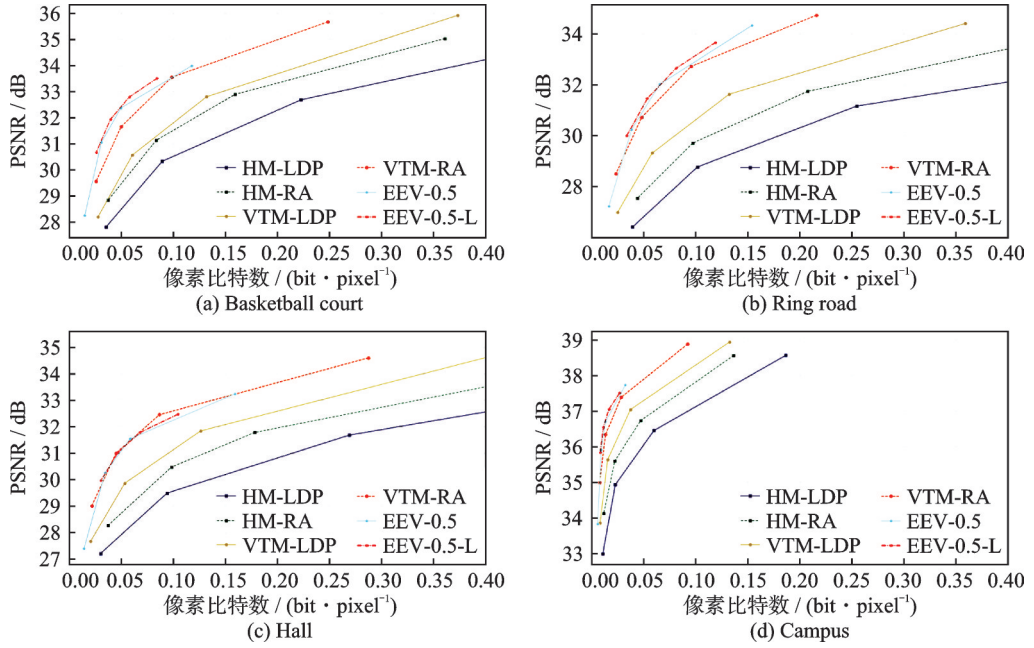


图 8 RGB-PSNR 指标下的码率-失真曲线

Fig.8 Rate-distortion curves for RGB-PSNR metrics

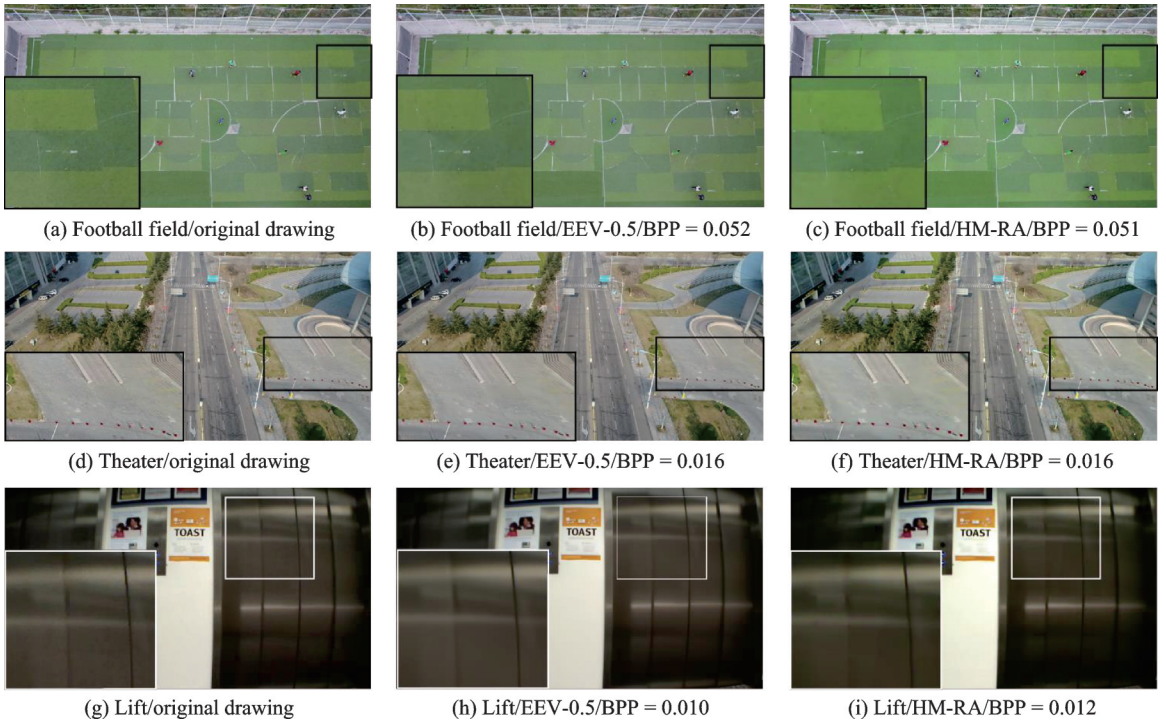


图 9 RGB-PSNR 指标下的主观质量对比图

Fig.9 Subjective quality comparison for RGB-PSNR metrics

对于 MS-SSIM 指标，EEV-0.5 的平均性能与 VTM-RA 相比仍存在约 3% 的差距，在 C 类测试集上的差距尤为明显。此时，EEV-0.5-L 表现出了较大的性能优势，BD-rate 指标相较于 EEV-0.5 实现了

7.12%的提升,超过了VTM-RA,实现了最优性能。

综上所述,EEV-0.5及其扩展版本EEV-0.5-L在PSNR和MS-SSIM指标上的优异表现,不仅验证了其在编码效率和质量上的显著提升,也展示了更大规模模型在复杂场景下的巨大潜力。未来,随着模型规模的进一步扩展和优化,EEV系列参考软件有望在视频编码领域实现更多突破,为高效视频压缩技术的发展提供重要支持。

3.4　分析与展望

无人机视频编码在神经网络编码领域具有重要的研究价值和应用潜力,随着无人机技术的快速发展和应用场景的不断扩展,无人机视频数据的复杂性和规模进一步增加,对编码技术也提出了更高的要求。虽然目前提出的算法已经实现了较好的平均编码性能,但在某些特定场景下,其性能与传统编码标准VTM-RA相比仍存在一定差距。这些差距主要体现在以下几个方面:

(1)高分辨率与大幅运动场景的挑战:首先,对于分辨率较大的立交桥(2 720像素×1 520像素)序列,EEV-0.5和EEV-0.5-L的编码表现均不佳。这可能是由于序列中存在着幅度较大的运动,而当前算法采用的光流网络难以对这些大幅运动进行准确估计,导致预测质量和残差编码效果不佳。

(2)室内复杂场景的优化:对于室内场景,虽然EEV-0.5-L相较于EEV-0.5有一定的性能提升,但是还存在着较大的优化空间。室内场景通常包含复杂的物体布局、较大的镜头畸变以及严重的遮挡问题,这些因素对编码算法提出了更高的要求。

(3)计算效率与实时性需求:虽然更大规模的模型(如EEV-0.5-L)能够带来更优的编码性能,但其计算量和编解码时间也显著增加。然而,无人机作为资源受限的计算平台,其算力和存储能力有限,难以支持大规模模型的高效运行。以Neousys宸曜科技的FLYC-300轻量级无人机任务计算平台为例,其搭载的NVIDIA® Jetson Orin™ NX可以提供100TOPs的算力,其显存大小支持本算法对1080P视频进行编解码。在此算力及FP32精度下,本文模型对于480P尺寸的视频,处理速度约4帧/s;对于720P尺寸的视频,处理速度约1.25帧/s;而对于1080P尺寸的视频,处理速度仅有0.3帧/s,难以满足实际需求。因此,未来的研究需要探索轻量化网络设计和高效计算架构,例如通过模型剪枝、量化、知识蒸馏等技术减少模型参数量,同时结合硬件加速(如图形处理器或专用人工智能芯片)以实现实时编码。

(4)数据集与评估标准的完善:当前针对无人机视频的公开数据集和评估标准相对有限,尤其是高质量的训练数据集,这在一定程度上限制了相关研究的进一步发展。未来可以构建更加多样化的无人机视频数据集,涵盖不同分辨率和帧率,以及更丰富的环境条件和特性,以更好地训练、测试和比较不同编码方法的性能。

为系统验证算法先进性,本节在无人机视频的基准测试集上进行了广泛测试,提出的算法在不同测试指标上均展示出良好的性能,证明了该框架的先进性和高效性。此外,对无人机视频编码的分析与展望也为该领域未来的探索指明了可行的发展方向。

4　结束语

本文提出一种基于分层双向运动模型的端到端视频编码算法,通过参数共享的运动编解码器架构、分层双向运动表示及可信运动建模机制,算法显著提升了视频压缩的效率。该算法在无人机视频处理中展现出显著优势,有效应对运动模糊、尺度变化、复杂背景干扰及镜头畸变等挑战,验证了复杂场景下的鲁棒压缩能力。研究同时为MPAI-EEV标准演进提供了层间码率分配策略优化和模型规模扩展的创新路径,并通过深度解析无人机视频编码的特殊性,拓展了端到端编码技术的理论框架与应用边界。随着人工智能技术的迭代升级,端到端视频编码模型有望在更多复杂场景中得到广泛应用,

为视频编码领域带来新的突破。

参考文献：

[1] 贾川民, 马海川, 杨文瀚,等. 视频处理与压缩技术[J]. 中国图象图形学报, 2021, 26(6): 1179-1200.
JIA Chuanmin, MA Haichuan, YANG Wenhan, et al. Video processing and compression technologies[J]. Journal of Image and Graphics, 2021, 26(6): 1179-1200.

[2] ZHU P, WANG L, DU D, et al. Detection and tracking meet drones challenge[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(11): 7380-7399.

[3] 田畅, 贾倩, 陈润丰, 等. 无人机集群网络资源优化综述[J]. 数据采集与处理, 2023, 38(3): 506-524.
TIAN Chang, JIA Qian, CHEN Runfeng, et al. Review on optimization of resources in UAV swarm networks[J]. Journal of Data Acquisition and Processing, 2023, 38(3): 506-524.

[4] RIPPEL O, ANDERSON A, TATWAWADI K, et al. ELF-VC: Efficient learned flexible-rate video coding[C]// Proceedings of IEEE International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021.

[5] LI J, LI B, LU Y. Hybrid spatial-temporal entropy modelling for neural video compression[C]//Proceedings of the ACM International Conference on Multimedia. Lisbon, Portugal: ACM, 2022.

[6] 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 (CVPR). Vancouver, Canada: IEEE, 2023.

[7] LI J, LI B, LU Y. Deep contextual video compression[C]//Proceedings of the Advances in Neural Information Processing Systems, Virtual Conference. San Diego, USA: NIPS, 2021: 18114-18125.

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

[9] 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.

[10] LIU H, LU M, MA Z, et al. Neural video coding using multiscale motion compensation and spatiotemporal context model[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(8): 3182-3196.

[11] HU Z, LU G, GUO J, et al. Coarse-to-fine deep video coding with hyperprior-guided mode prediction[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022.

[12] 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, 2024, 34(5): 3096-3110.

[13] 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 (CVPR). Seattle, USA: IEEE, 2020.

[14] 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. Nashville, USA: IEEE, 2021: 1502-1511.

[15] FENG R, GUO Z, ZHANG Z, et al. Versatile learned video compression[EB/OL]. (2021-11-05)[2025-02-20]. https://arxiv.org/abs/2111.03386.

[16] HO Y, CHANG C, CHEN P, et al. CANF-VC: Conditional augmented normalizing flows for video compression[C]// Proceedings of the European Conference on Computer Vision. Berlin, Germany: Springer, 2022: 207-223.

[17] CHEN T, LIU H, SHEN Q, et al. DeepCoder: A deep neural network based video compression[C]//Proceedings of the IEEE Visual Communications and Image Processing. St. Petersburg, USA: IEEE, 2017.

[18] 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. Long Beach, USA: IEEE, 2019: 10998-11007.

[19] HU Z, CHEN Z, XU D, et al. Improving deep video compression by resolution-adaptive flow coding[C]//Proceedings of the European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020.

[20] LU G, ZHANG X, OUYANG W, et al. An end-to-end learning framework for video compression[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3292-3308.

[21] YANG R, MENTZER F, VAN GOOL L, et al. Learning for video compression with recurrent auto-encoder and recurrent probability model[J]. IEEE Journal of Selected Topics in Signal Processing, 2021, 15(2): 388-401.

[22] SHI Y, GE Y, WANG J, et al. AlphaVC: High-performance and efficient learned video compression[C]//Proceedings of the European Conference on Computer Vision (ECCV). Tel Aviv, Israel: Springer, 2022.
[23] WU C, SINGHAL N, KRÄHENBÜHL P. Video compression through image interpolation[C]//Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer, 2018: 416-431.
[24] DJELOUAH A, CAMPOS J, SCHAUB-MEYER S, et al. Neural inter-frame compression for video coding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019: 6420-6428.
[25] POURREZA R, CHEN T. Extending neural P-frame codecs for B-frame coding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021: 6680-6689.
[26] ALEXANDRE D, HANG H, PENG W. Hierarchical B-frame video coding using two-layer CANF without motion coding [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023: 10249-10258.
[27] YANG R, MA M, GU L, et al. Learning for video compression with hierarchical quality and recurrent enhancement[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020: 6628-6637.
[28] KIM Y, SEO B, LEE W, et al. Neural video compression with temporal layer-adaptive hierarchical B-frame coding[EB/OL]. (2023-08-30)[2024-02-20]. https://arxiv.org/abs/2308.15791.
[29] YE F, LI Z, CHEN J. Deep video compression with scaled hierarchical bi-directional motion model[C]//Proceedings of the ACM International Conference on Multimedia. Melbourne, Australia: ACM, 2024: 11244-11247.
[30] SHENG X, LI L, DENG L, et al. Bi-directional deep contextual video compression[EB/OL]. (2024-08-08)[2025-02-20]. https://arxiv.org/abs/2408.08604.
[31] FENG R, WU Y, GUO Z, et al. Learned video compression with feature-level residuals[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA: IEEE, 2020: 529-532.
[32] LADUNE T, PHILIPPE P, HAMIDOUCHE W, et al. Optical flow and mode selection for learning-based video coding[C]//Proceedings of the IEEE 22nd International Workshop on Multimedia Signal Processing. Tampere, Finland: IEEE, 2020: 1-6.
[33] BELYAEV E, SKOBLEV F. An efficient storage of infrared video of drone inspections via iterative aerial map construction[J]. IEEE Signal Processing Letters, 2019, 26(8): 1157-1161.
[34] BHASKARANAND M, GIBSON J D. Global motion assisted low complexity video encoding for UAV applications[J]. IEEE Journal of Selected Topics in Signal Processing, 2014, 9(1): 139-150.
[35] JIANG W, WANG R. MLIC++: Linear complexity multi-reference entropy modeling for learned image compression[C]//Proceedings of the ICML 2023 Workshop on Neural Compression. Honolulu, USA: ACM, 2023.
[36] JIA C, YE F, SUN H, et al. Learning to compress unmanned aerial vehicle(UAV) captured video: Benchmark and analysis [C]//Proceedings of the Data Compression Conference. Snowbird, USA: IEEE, 2023.
[37] KUMAR A, BOUGANIS C. Learning to fly by myself: A self-supervised CNN-based approach for autonomous navigation [C]//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Madrid, Spain: IEEE, 2018: 5216-5223.
[38] DU D, YU Q, HUA Y, et al. The unmanned aerial vehicle benchmark: Object detection and tracking[C]//Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer, 2018: 370-386.

作者简介:

叶枫(2002-):女,博士研究生,研究方向:端到端视频编码,E-mail:feng.ye@stu.pku.edu.cn。

董凡可(2002-),男,博士研究生,研究方向:图像视频无损编码。

贾川民(1993-),**通信作者**:男,助理教授,研究方向:智能视频编码率失真理论与高效编码方法,E-mail:cmjia@pku.edu.cn。

(编辑:张蓓,王婕)