ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 2,Mar. 2025,pp. 320－333
DOI:10. 16337/j. 1004-9037. 2025. 02. 004

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

低码率生成式无人机视频编码算法

刘美琴[1,2]，陈虹宇[1,2]，周一鸣[1,2]，倪文昊[1,2]

(1. 北京交通大学信息科学研究所,北京 100044;2. 北京交通大学视觉智能交叉创新教育部国际合作联合实验室,北京 100044)

摘　要：空天地海复杂环境下海量的视频数据给有限的传输带宽和存储设备带来了巨大的压力,因此如何提高视频编码技术在低码率条件下的编码效率显得尤为关键。近年来,基于深度学习的视频编码算法取得了良好的进展,却因优化目标与感知质量失配、训练数据分布偏差等问题,降低了极低码率下的视觉感知质量。生成式编码通过学习数据分布有效提升了低码率下的纹理与结构复原能力,缓解了深度视频压缩的模糊伪影问题。然而,现有研究仍存在两大瓶颈:一是时域相关性建模不足,帧间关联缺失;二是动态比特分配机制欠缺,难以实现关键信息的自适应提取。为此,提出一种基于条件引导扩散模型的视频编码算法(Conditional guided diffusion model-video compression, CGDM-VC),旨在改善低码率条件下视频感知质量的同时,加强帧间特征建模能力和保留关键信息。具体地,该算法设计了隐式帧间对齐策略,利用扩散模型捕获帧间潜在特征,降低估计显式运动信息的计算复杂度。同时,设计的自适应时空重要性编码器可动态分配码率优化关键区域的生成质量。此外,引入感知损失函数,结合感知图像块相似度(Learned perceptual image patch similarity, LPIPS)约束,以提高重建帧的视觉保真度。实验结果表明,与DCVC(Deep contextual video compression)等算法相比,该算法在低码率(＜0.1 BPP)情况下,LPIPS值平均降低了36.49%,展现出更丰富的纹理细节和更自然的视觉效果。

关键词：视频编码;扩散模型;感知质量;帧间对齐;低码率

中图分类号：TP391　　**文献标志码**:A

Low Bit Rate Generative Drone Video Compression

LIU Meiqin[1,2], CHEN Hongyu[1,2], ZHOU Yiming[1,2], NI Wenhao[1,2]

(1. Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China; 2. Visual Intelligence＋X International Cooperation Joint Laboratory, Beijing Jiaotong University, Beijing 100044, China)

Abstract: In complex environments across air, space, land, and sea, the massive volume of video data exerts tremendous pressure on limited transmission bandwidth and storage devices. Therefore, improving the coding efficiency of video compression technologies under low bit rate conditions becomes crucial. In recent years, deep learning-based video compression algorithms have made significant progress, yet due to issues such as model design flaws, mismatches between optimization objectives and perceptual quality, and biases in training data distributions, the visual perception quality at extremely low bit rates has been compromised. Generative encoding effectively improves the texture and structure restoration ability at low bit rates through data distribution learning, alleviating the problem of blur artifacts in deep video

基金项目:国家自然科学基金(62372036)。
收稿日期:2025-02-15;**修订日期**:2025-03-14

compression. However, there are still two major bottlenecks in existing research: Firstly, time domain correlation modeling is insufficient and inter-frame feature correlation is missing; secondly, the lack of dynamic bit allocation mechanism makes it difficult to achieve adaptive extraction of key information. Therefore, this article proposes a video encoding algorithm based on conditional guided diffusion model-video compression (CGDM-VC), aiming to improve the perceptual quality of videos under low bit-rate conditions while enhancing inter-frame feature modeling capabilities and preserving key information. Specifically, the algorithm designs an implicit inter-frame alignment strategy, utilizing a diffusion model to capture potential inter-frame features and reduce the computational complexity of estimating explicit motion information. Meanwhile, the designed adaptive spatio-temporal importance-aware coder can dynamically allocate code rates to optimize the generation quality of key regions. Furthermore, a perceptual loss function is introduced, combined with the learned perceptual image patch similarity (LPIPS) constraint, to improve the visual fidelity of the reconstructed frames. Experimental results demonstrate that, compared to algorithms such as deep contextual video compression (DCVC), the proposed method achieves an average LPIPS reduction of 36.49% under low bit rate conditions (<0.1 BPP), showing richer texture details and more natural visual effects.

Key words: video compression; diffusion model; perceptual quality; inter-frame alignment; low bit rate

引　　言

随着航空航天、遥感监测及通信技术的不断发展,海量的视频数据面临着巨大的存储与传输压力,尤其在空天地海等特定场景中,视频数据的传输带宽常常受到严格限制[1]。在高分辨率和高帧率视频应用场景下,传统视频存储和传输系统难以满足低带宽和高效能的需求[2-3],因此如何在有限的存储和传输资源下保持低码率条件下视频数据的感知质量,已成为当前空天地海视觉信息智能处理领域的重要研究课题之一。传统视频编码标准,如H.264/AVC[4]和H.265/HEVC[5],利用运动估计与运动补偿机制去除视频帧内的空间冗余和帧间的时间冗余,缓解了空天地海场景中一定的视频存储和传输的压力。然而,码率越低,重建视频的块效应现象越明显,严重影响了视频的视觉质量。目前,视频编码框架主要依赖于手工设计的复杂变换和编码策略,难以适应不同场景和内容的动态特性,尤其是高压缩比和低码率环境。近年来,基于神经网络的视频编码算法因其能够自动学习视频的内在特征而逐渐成为研究热点。这些算法克服了传统编码方法中手工设计策略的局限性,能够更好地适应视频内容的复杂性和场景的多样性,从而在低码率环境中展现出更好的压缩性能和感知质量[6]。

随着越来越多的新技术引入,基于神经网络的视频编码算法呈现出多样化的发展趋势,主要分为3类。具体地,基于神经网络的视频编码算法主要包括基于传统神经网络、生成对抗网络(Generative adversarial networks, GAN)以及扩散模型的视频编码算法[7]。其中,基于传统神经网络的深度视频编码算法[8-15]已有较长的研究历史,基于GAN的深度视频编码算法[16-21]则是近年来新兴研究方向,基于扩散模型的深度视频编码算法[22-23]亦刚崭露头角。基于传统神经网络的视频编码通常利用神经网络替代传统视频编码中的各个模块(如运动估计、变换、量化和熵编码等),并通过端到端的训练方式实现编码和解码优化。然而,在低码率条件下,显式光流估计难以维持较高的重建精度,导致重建视频帧出现明显的运动模糊和压缩伪影,降低了视频的视觉感知质量。基于GAN的视频编码则利用生成对抗网络的生成能力,获得高质量的重建图像,可在低比特率条件下平衡重建质量和压缩性能,但其训练过程会不稳定且易出现模式崩溃问题[24-25]。基于扩散模型的编码算法通过模拟视频数据的逐步去噪过程,能

够在低比特率下保留更多的纹理和结构信息,生成高质量的重建结果,具备较强的鲁棒性。同时,该算法避免了传统编码框架对显式运动补偿的依赖,可更好地适用于运动复杂的场景中。然而,扩散模型计算复杂度高、帧间一致性利用率低,直接调用扩散模型进行视频编码难以满足空天地海场景的实时传输需求。

为此,本文提出了一种基于条件引导扩散模型的视频编码算法(Conditional guided diffusion model-video compression, CGDM-VC)。为了增强帧间特征的建模能力,提出了隐式帧间对齐策略(Implicit inter-frame alignment strategy, IIAS)。该策略以隐式扩散模型为基础,通过引入特征对齐优化机制,显著改善了帧间特征的有效利用。同时,设计了自适应时空重要性编码器(Adaptive spatio-temporal importance-aware coder, ASTC),利用时空注意力网络从参考特征中提取时空重要性权重,用于缩放当前目标特征并进行码率分配,确保了编码器能够根据时空特征的重要性动态调整码率,提高关键区域的压缩质量,同时减少次要区域的比特开销。具体地,针对低码率传输场景下视觉感知质量劣化的瓶颈问题,创新性地提出潜空间映射方法,将视频压缩过程迁移至潜空间域进行特征重构。该方法通过构建隐式特征扩散网络,在避免显式光流估计带来运动估计误差累积的同时,有效消除传统方法中光流辅助分支的压缩损耗,在抑制压缩伪影与降低整体码率开销方面展现双重技术优势。此外,设计了自适应时空重要性编码器,并提出时空调制单元(Spatio-temporal modulation unit, STMU)对目标特征进行特征缩放,借助神经网络提取每一帧的重要性权重,根据运动复杂度、纹理复杂度等因素动态调整码率分配,从而提升重建质量。同时,通过结合学习感知图像块相似度[26],优化了低码率环境下的视频视觉质量,突破了传统评价方法的局限性。实验结果表明,在无人机数据集(UAVDT[27]、ERA[28]、AU-AIR[29])和传统数据集(HEVC[30])上,与DCVC(Deep contextual video compression)等视频编码算法[9-13]相比,本文算法不仅在低码率条件下可显著提高视频的重建质量,而且可获得丰富的纹理细节和良好的压缩效率,适用于空天地海等特定应用场景中。

本文的主要贡献如下:(1) 设计了基于条件引导的生成式无人机视频编码算法,该算法创新性地将扩散模型的条件引导机制与帧间关系建模相结合,充分利用扩散模型的先验信息,实现了低码率条件下高感知质量的视频重建;(2) 设计了隐式帧间对齐策略,通过构建特征对齐优化机制,避免了低码率条件下显式运动估计带来的模糊伪影问题,同时节省了运动信息的码率开销,有效解决了显式运动建模的固有缺陷;(3) 设计了自适应时空重要性模块,通过时空注意力网络捕捉视频帧间潜在特征的时空关联性,动态调整信息编码策略,分配码率优化关键区域的重建质量,降低了码率开销,提升了整体视频的感知质量。

1 相关工作

根据使用神经网络的不同,基于深度视频编码算法可分为3类:基于传统神经网络的深度视频编码算法、基于生成对抗网络的编码算法和基于扩散模型的编码算法。

1.1 基于传统神经网络的深度视频编码算法

基于传统神经网络的深度视频编码算法通过神经网络自动学习视频数据的内在特征,利用端到端的训练方式替代了传统视频编码中的多个模块,如运动估计、变换、量化和熵编码等,以提升视频编码的性能和压缩效率。与传统编码算法相比,基于传统神经网络的深度视频编码算法能够自动适应视频内容的复杂性,从而在不同场景下提供更加灵活和高效的视频编码方案。如,Lu等[8]提出了首个端对端的基于深度学习的视频编码算法(Deep video compression, DVC),该方法通过将传统编码框架中的各个模块替换为神经网络实现视频的高效压缩和重建。Li等[9]提出了一种基于上下文的视频编码框架DCVC,旨在实现从残差编码到条件编码的范式转换。研究者们又设计DCVC的多个衍生模型,如DCVC-TCM[10]专注于挖掘时间上下文信息,通过学习视频编码中的时间相关性,进一步提高编码性

能;DCVC-HEM[11]引入了混合时空熵建模技术学习空间相关性;DCVC-DC[12]引入了多样化的上下文信息进一步利用空间相关性,并提出了棋盘格熵编码,提高熵建模的准确性和压缩效率;DCVC-FM[13]采用特征调制方法,实现了仅使用一个I帧的端到端神经视频编解码器,并扩展了码率,实现了全码率压缩;Sheng等[14]尝试在基线模型的基础上构建大型神经视频编码模型,显著提升了视频压缩性能;Sheng等[15]结合CNN和时序建模,优化帧内/帧间预测模块,并在特征域完成运动补偿和跨域编码。

1.2 基于生成对抗网络的深度视频编码算法

基于生成对抗网络的深度视频编码算法利用GAN网络的生成能力来提高视频压缩的效果。该网络由生成器和判别器构成,生成器负责生成视频帧,判别器则用于评估生成视频帧的质量。在视频编码任务中,生成器通过学习视频数据的潜在分布生成逼真的视频帧,判别器则用于确保生成帧在视觉上接近真实内容[31]。这种生成对抗的方式能够有效地减少块效应和伪影,尤其在低比特率视频重建中表现突出。如,Hu等[16]提出了一种面向人类视觉和机器视觉的可扩展图像编码框架,利用GAN从紧凑的特征表示和参考像素重建图像,在人眼视觉质量和面部特征检测方面都取得优异的效果;Chang等[17]提出了一种分层融合GAN(Hierarchical fusion GAN, HF-GAN),将视觉数据分解为紧凑的结构和纹理表示,而不是去除信号级的冗余信息;Mentzer等[18]提出了一种基于GAN的视频编码算法,通过合成和传播细节,利用高质量的光流提升了视觉质量;Dundar等[19]通过GAN生成多视角图像数据集,优化3D重建中的跨视角一致性,间接提升视频动态场景编码效率;Lan等[20]提出了一种多视图视频编码算法,利用GAN网络的生成能力实现中间视点的精确重建。

1.3 基于扩散模型的深度视频编码算法

近年来,扩散模型在生成领域发展迅速。如,Ho等[21]提出的DDPM (Denoising diffusion probabilistic models)通过逐步去噪的方式生成图像,取得了比传统生成模型更优的生成效果,特别适用于需要重建高质量视频的编码任务。然而,DDPM的计算复杂度非常高,限制了其在实时视频编码中的应用。为此,Rombach等[22]提出LDM (Latent diffusion models),在潜在空间进行扩散,在保持生成视频帧细节信息的同时,降低了计算复杂度。在低比特率视频编码领域,Voleti等[23]提出了一种多功能视频生成模型MCVD(Masked conditional video diffusion),具备视频生成、预测和重建的能力,在多种视频处理任务中表现出色。虽然基于扩散模型的视频编码算法表现突出,却仍然面临计算效率和实时处理的挑战。因此,研究者们在扩散模型中引入了多种优化策略,如加速生成过程、引入并行处理等,旨在提高扩散模型在视频任务中的应用效率。如,Li等[32]提出了DiffVSR框架,优化生成路径将逐帧采样方式改进为时序感知的联合优化,并通过多尺度时间注意力模块捕捉长程运动的依赖关系,加速了生成过程;Mao等[33]提出了OSV模型,设计了潜在空间分布式采样策略,通过并行生成和对抗训练加快生成速度;Li等[34]提出了一种集成迁移攻击与查询攻击的框架,优化了视频编码中的对抗样本生成与防御。

2 算法框架

2.1 条件引导扩散模型的视频编码算法

为了提高低码率条件下的视频的感知质量,本文提出了一种基于条件引导扩散模型的视频编码算法CGDM-VC,结构如图1所示。该算法引入了时空重要性引导的可变码率编解码器,将参考帧的特征$\hat{y}_{\text{ref}}$作为先验信息,与目标特征y_i融合后得到解码特征$\hat{y}_i$,避免了借助显式光流或显式

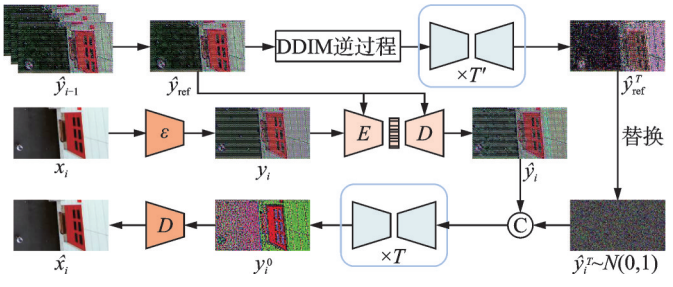

图1 条件引导扩散模型的视频编码算法示意图

Fig.1 Framework of CGDM-VC

运动信息，消除了帧间冗余。同时，设计了基于扩散模型的帧间对齐策略，通过去除帧间运动信息的显式残差信息，提高了低码率下的编码效率，并引入解码特征 $\hat{y}_i$ 作为从初始特征 y_i^T 到目标特征 y_i 的最大化似然估计条件，以保证视频帧的重建质量。

2.2 自适应时空重要性编码器

在传统的视频编码中，编码显式的运动信息是优化重建质量的重要手段之一，通过捕捉帧间的运动变化来消除时间冗余。然而，传输运动信息所需的开销给低码率视频编码提出了严峻挑战。若不采用编码显式运动信息的方式会带来帧间信息的缺失。因此，为了有效捕捉视频帧间潜在的时空关联性，提出了一种自适应时空重要性编码器，结构如图2所示。该编码器通过建模视频帧间的时空特性，动态调整信息编码策略，以优化视频压缩与重建过程。ASTC将参考帧特征 $\hat{y}_{\text{ref}}$ 作为先验信息编码目标特征 y_i 得到解码特征 $\hat{y}_i$，可表示为

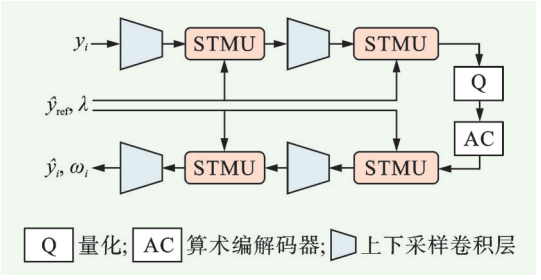

图2 自适应时空重要性编码器

Fig.2 Adaptive spatio-temporal importance-aware coder

$$\hat{y}_i = D(\lfloor E(y_i, \hat{y}_{\text{ref}}) \rceil, \hat{y}_{\text{ref}}) \tag{1}$$

式中：$E(\cdot)$ 和 $D(\cdot)$ 分别表示具有4倍上下采样卷积层和两个STMU的编码器和解码器。

在传统编码器中，不同区域的编码策略是固定的，本文设计的STMU利用神经网络提取每一帧的重要性权重，根据运动复杂度、纹理复杂度等因素动态调整码率分配，结构如图3所示。

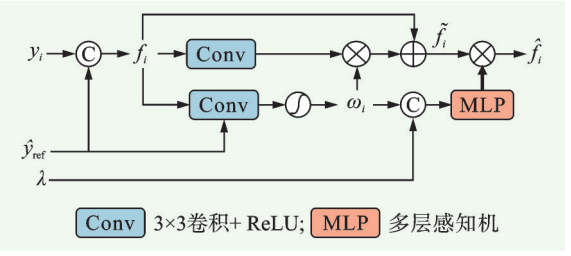

图3 时空调制单元

Fig.3 Spatio-temporal modulation unit

具体地，STMU利用时空注意力网络从参考特征 $\hat{y}_{\text{ref}}$ 中提取时空重要性权重 ω_i，用于缩放当前目标特征 y_i 并进行码率分配，可表示为

$$\omega_i = \sigma\left(\text{Conv}(\text{Concat}(f_i, \hat{y}_{\text{ref}}))\right) \tag{2}$$

式中：f_i 表示 $\hat{y}_{\text{ref}}$ 和 y_i 的融合特征，大小为(6,112,208)；Concat(·)表示融合操作，以batch=1为例，f_i 和 $\hat{y}_{\text{ref}}$ 的大小分别为(3,112,208)、(3,112,208)；Conv(·)表示1个具有ReLU激活函数的3×3卷积层；$\sigma(\cdot)$ 表示Sigmoid函数，用于归一化注意力权重，使其值限定为[0,1]。接着，STMU利用权重 ω_i 提取目标特征 y_i 中的时空特征 $\tilde{f}_i$，可表示为

$$\tilde{f}_i = f_i + \omega_i \cdot \text{Conv}(f_i) \tag{3}$$

式中：ω_i 表示时空重要性权重，Conv(·)表示1个具有ReLU激活函数的3×3卷积层。STMU融合时空注意力信息后，通过 ω_i 赋予目标特征不同的重要性，增强关键区域的特征，抑制了非关键区域的特征。此外，该模块可自适应地调整特征表达方式，通过残差连接，确保特征调整不会导致信息损失，提高了模型的稳定性。STMU进一步结合重要性权重 ω_i 和码率权重 λ 动态分配码率，得到缩放后的特征 $\hat{f}_i$，大小为(192,28,52)，可表示为

$$\hat{f}_i = \text{MLP}(\text{Concat}(\omega_i, \lambda)) \cdot \tilde{f}_i \tag{4}$$

式中：λ 表示码率控制权重，决定目标特征在不同区域的比特分配；$\tilde{f}_i$ 表示时空特征，大小为(192,56,104)；Concat(·)表示融合操作，Concat(ω_i, λ)结合了时空注意力权重和码率控制因素，自适应地分配码

率,以batch=1为例,ω_i和λ的大小分别为(4,1)和(4,1);MLP(•)表示多层感知机,用于学习码率控制策略。这一过程确保了编码器能够根据时空特征的重要性动态调整码率,提高关键区域的压缩质量,同时减少次要区域的比特开销。

图4呈现了经ASTC量化后的细节特征码率分布热力图。在视频帧中,诸如马路街道以及等待红绿灯的车辆等相对静止的区域,在视频序列中较为常见且变化幅度较小,ASTC能够高效地为这些区域分配较少的比特进行表示。尽管这些区域可能包含一定量的细节信息,但只需较低的码率即可完成编码。相对而言,对于车辆运动轨迹等具有独特性且动态变化显著的对象,ASTC会自适应地为其分配更多比特以进行编码。这表明ASTC能够依据视频帧间的潜在时空关联性,动态调整信息编码策略。在确保关键区域信息质量不受损失的前提下,有效降低了整体码率,从而实现了更为高效的视频压缩与重建。

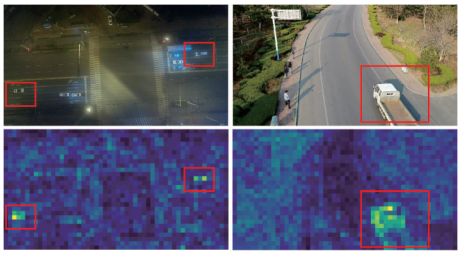

图4 量化细节特征的熵分布

Fig.4 Quantifying the entropy distribution of detailed features

2.3 隐式帧间对齐策略

在传统的视频编码方法中,显式的帧间对齐,如运动估计和运动补偿,是提升视频重建质量的关键技术。然而,在低码率条件下,去除显式运动信息编码可能会影响帧间对齐。显式帧间对齐依赖于复杂的运动估计方法,在高分辨率视频或长时间序列中需要大量的计算开销。因此,为了在不显式编码运动信息的情况下捕捉和利用帧间信息,本文提出隐式帧间对齐策略,利用前一帧的信息,通过U-Net学习帧间的时空关联性,以实现隐式的帧间对齐。

具体而言,提取前一帧的特征后与当前帧的信息进行融合,通过U-Net网络的结构将前一帧的信息转化与当前帧对齐以捕捉帧间潜在的时空关系。这一过程不仅能够有效减少传统运动补偿中产生的冗余信息,还能在低码率条件下保留更多重要的细节和结构信息。为此,设计了隐式帧间对齐模块,该模块在隐式扩散模型基础上优化帧间特征的对齐过程,更好地保留和传递视频的时间信息。隐式扩散模型结构如图5所示,主要包括编码器、U-Net和解码器3部分。输入数据通过编码器映射到隐空间,并进行前向加噪操作得到z_T,接着提取条件τ_θ,与加噪后的隐空间表示融合,输入U-Net网络进行反向去噪后,完成解码过程。

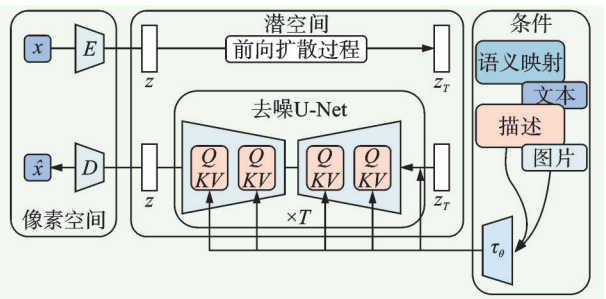

图5 隐式扩散模型

Fig.5 Implicit diffusion models

在CGDM-VC的解码过程中,解码特征$\hat{y}_i$作为从初始特征y_i^T到目标特征y_i的最大化似然估计条件,通过隐式扩散模型进行帧间对齐,通过$p_\theta(y_i^{0:T}|y_i)$求解与y_i维数相等的隐变量$y_i^{0:T}$,可表示为

$$p_\theta(y_i^{0:T}|y_i)=p(y_i^T)\prod_{t=1}^{T}p_\theta(y_i^{t-1}|y_i^t,\hat{y}_i) \tag{5}$$

式中:$p(y_i^T)=\mathcal{N}\left(\hat{y}_{\text{ref}}^{T'};\mu_\theta(\hat{y}_{\text{ref}}^{T'},T'),\sigma_\theta^2(\hat{y}_{\text{ref}}^{T'},T')\right)$表示最大化似然估计初始状态,$\hat{y}_{\text{ref}}^{T'}$表示对参考特征$\hat{y}_{\text{ref}}$进行DDIM逆变换的结果,$T'=\frac{1}{2}T$表示DDIM逆变换的步数,$t=T\rightarrow 1(T=30)$表示求解最大似然

估计的步数。每一步状态转移过程$p_\theta(y_i^{t-1}|y_i^t,\hat{y}_i)$通过LDM[22]预训练U-Net网络进行计算,可表示为

$$p_\theta(y_i^{t-1}|y_i^t,\hat{y}_i)=N\left(y_i^{t-1};\mu_\theta(y_i^t,t,\hat{y}_i),\sigma_\theta^2(y_i^t,t,\hat{y}_i)\right) \tag{6}$$

式中:$\mu_\theta(y_i^t,t,\hat{y}_i)$和$\sigma_\theta^2(y_i^t,t,\hat{y}_i)$分别表示状态转移过程中拟合的均值和方差。由于解码特征$\hat{y}_i$与初始特征$\hat{y}_{\text{ref}}^{T'}$在结构和语义上具有相似性,时间域的运动信息有所不同,因此该状态转移过程相比式(5),可更准确地预测每步的均值μ_θ和方差σ_θ^2。

为了在减少编码信息量和不编码运动信息的同时,能保证视频帧的重建质量,隐式帧间对齐模块,以解码特征$\hat{y}_i$作为预训练扩散模型——LDM图像超分辨率模型[22]的条件,恢复目标特征y_i,重建视频帧,结构如图6所示。CGDM-VC采用隐式帧间对齐策略,利用参考帧的潜空间信息和运动信息保证重建质量,消除LDM预训练模型在训练时未考虑到帧间参考信息,且最大化似然函数过程的初始分布为随机噪声$\varepsilon\sim N(0,1)$。

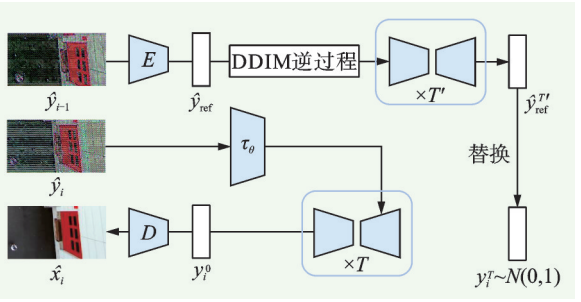

图6 隐式帧间对齐模块

Fig.6 Implicit inter-frame feature alignment module

为了在扩散过程中有效对齐帧间信息,CGDM-VC首先对参考特征$\hat{y}_{\text{ref}}$进行DDIM逆扩散加噪使其接近于高斯噪声的分布,用$\hat{y}_{\text{ref}}^{T'}$来替换原本的随机噪声$\varepsilon\sim N(0,1)$作为扩散模型的初始输入。求解$\hat{y}_{\text{ref}}^{T'}$需使用由$\hat{y}_{\text{ref}}^{t-1}$到$\hat{y}_{\text{ref}}^{t}$的递推公式,可表示为

$$\hat{y}_{\text{ref}}^{t}=\sqrt{\alpha^t}\,\frac{\hat{y}_{\text{ref}}^{t-1}-\sqrt{1-\alpha^{t-1}}\,\omega_i\varepsilon_\theta}{\sqrt{\alpha^{t-1}}}+\sqrt{1-\alpha^t}\,\omega_i\varepsilon_\theta \tag{7}$$

式中:$\hat{y}_{\text{ref}}^{0}=\hat{y}_{\text{ref}}$表示初始参考特征,直接从前一帧提取的特征,$\alpha^{t-1}$和$\alpha^t$表示扩散模型的加噪权重系数,$\varepsilon_\theta$表示扩散噪声预测模型的输出。该过程可生成一个更加符合视频帧分布的初始化特征$\hat{y}_{\text{ref}}^{T'}$,减小直接使用高斯噪声作为初始输入带来的分布偏差。在编码端,CGDM-VC对目标特征$\hat{y}_i$与初始化特征$\hat{y}_{\text{ref}}^{T'}$进行帧间对齐,并利用U-Net进行去噪优化,以得到高质量的潜空间特征y_i^0。扩散模型的状态转移过程$p_\theta(y_i^{t-1}|y_i^t)$可视为从初始特征$\hat{y}_{\text{ref}}^{T'}$逐步去噪至最大似然特征$y_i^0$的过程,可表示为

$$y_i^0=\hat{y}_{\text{ref}}^{T'}-\sum_{t=1}^{T}\varepsilon_\theta(y_i^t,\ t,\ \hat{y}_i) \tag{8}$$

式中:$\varepsilon_\theta(y_i^t,\ t,\ \hat{y}_i)$表示基于U-Net的去噪预测模型,为扩散去噪的总过程,t表示扩散步数。这一过程中,初始特征$\hat{y}_{\text{ref}}^{T'}$逐步去噪,最终得到高质量的潜空间特征$y_i^0$,可用于生成最终的解码视频帧。

从上述帧间对齐过程可看出,参考特征$\hat{y}_{\text{ref}}$在去噪过程中提供了目标特征y_i^0的纹理信息与结构信息,因此合理设置DDIM逆扩散步数T'对于帧间运动建模至关重要。假设扩散过程中的初始特征$\hat{y}_{\text{ref}}^{T'}$可表示为参考特征$\hat{y}_{\text{ref}}$与噪声$\sum_{t=1}^{T'}\varepsilon_\theta(\hat{y}_{\text{ref}},t)$之和,即

$$y_i^0=\hat{y}_{\text{ref}}+\sum_{t=1}^{T'}\varepsilon_\theta(\hat{y}_{\text{ref}},t)-\sum_{t=1}^{T}\varepsilon_\theta(y_i^t,\ t,\ \hat{y}_i) \tag{9}$$

式中:从参考特征$\hat{y}_{\text{ref}}$到最大似然特征y_i^0之间的运动信息$m_{\hat{y}_{\text{ref}}\to y_i^0}$可表示为

$$m_{\hat{y}_{\text{ref}}\to y_i^0}\approx y_i^0-\hat{y}_{\text{ref}}=\sum_{t=1}^{T'}\varepsilon_\theta(\hat{y}_{\text{ref}},t)-\sum_{t=1}^{T}\varepsilon_\theta \tag{10}$$

由式(10)可知,在扩散过程中,帧间运动信息的变化是由噪声估计模型ε_θ进行建模,从而避免了显

式的运动估计过程,实现了帧间信息的隐式建模。

参考Wang等[35]提出的方法,本文选择DDIM逆变换步数$T'=\frac{1}{2}T$以优化帧间预测质量。若T'过大,目标特征y_i与参考特征$\hat{y}_{\text{ref}}$之间的运动信息差距增大,模型需要额外的运动补偿信息,导致计算开销更大。若T'过小,目标特征y_i与参考特征$\hat{y}_{\text{ref}}$的运动信息接近,模型可更多地依赖$\hat{y}_{\text{ref}}$来减少编码条件$\hat{y}_i$的码率,提高压缩率。为了平衡帧间信息建模效果和码率分配之间的关系,本文设置DDIM逆变换步数为扩散过程步数的一半。

2.4 损失函数

本文设计了端到端的误差传播策略,在状态转移过程中,U-Net网络的参数保持固定不变,率失真损失函数$\mathcal{L}$可表示为

$$\mathcal{L}=\mathcal{R}(y_i)+\lambda\cdot(\mathcal{D}(x_i,\hat{x}_i)+\beta\cdot\mathcal{P}(x_i,\hat{x}_i)) \tag{11}$$

式中:λ表示码率$\mathcal{R}$和重建失真之间的平衡系数,β表示均方误差MSE损失$\mathcal{D}$和感知图像块相似度LPIPS[26]损失$\mathcal{P}$之间的权重系数。由于编码器包含可以调节特征选择的可变码率层,CGDM-VC仅需训练一个模型即可完成不同码率下的测试任务。

CGDM-VC算法流程如算法1所示。

算法1 条件引导扩散模型的视频编码算法

输入:x_i, $\hat{y}_{i-1}$ // 输入帧和特征缓存中参考特征

输出:$\hat{x}_i$, $\hat{y}_i$ // 输出帧和解码特征

(1) $y_i=\varepsilon(x_i)$ // 转换到特征域

(2) // 选择参考特征$\hat{y}_{\text{ref}}$和初始状态y_i^T

(3) $\hat{y}_{\text{ref}}\Leftarrow\hat{y}_{i-1}$ // 前向参考特征

(4) $y_i^T\Leftarrow\text{DDIM\_Inversion}(\hat{y}_{\text{ref}},\ T',\text{U-Net})$ // DDIM逆变换

(5) // 自适应时空重要性编码器

(6) $\hat{y}_i\Leftarrow D(\lfloor E(y_i,\hat{y}_{\text{ref}})\rceil,\hat{y}_{\text{ref}})$

(7) // 隐式帧间对齐

(8) for $t=T,T-1,\cdots,1$ do

(9) $y_i^{t-1}\Leftarrow\text{DDIM\_Backward}(y_i^t,t,\hat{y}_i,\text{U-Net})$ // 状态转移过程

(10) end for

(11) $\hat{x}_i=\mathcal{D}(y_i^0)$ // 帧重建

(12) return $\hat{x}_i$, $\hat{y}_i$

3 实 验

3.1 实验说明

训练集:采用常用的Vimeo-90K Septuplet[36]数据集训练编码模型。该数据集包含89 800个视频剪辑,覆盖了空天地海各种场景下的运动情况,每个视频序列由7个连续帧组成。经过预处理,实际使用了该数据集中的64 612个分辨率为256×256的7帧序列。

测试集:为了评估不同算法的性能,选取了UAVDT[27](forest、night、road)、ERA[28](Cycling、Fire、Harvesting)和AU-AIR[29]这3种经典的空天数据集,分辨率分别为1 280×704、640×640、832×448,视频序列帧数均为32。同时,也选取了传统的HEVC[30]数据集,包括Class-C(BasketballDrill、BQMall、Par-

tyScene、RaceHorses）、Class-D（BasketballPass、BlowingBubbles、BQSquare、RaceHorses）、Class-E（FourPeople、Johnny、KristenAndSara），分辨率分别为832×448、384×192、1 280×704，视频序列帧数均为32。

实验设置：实验在Intel(R) Xeon(R) Gold 6426Y CPU平台上完成，训练阶段采用4张NVIDIA GeForce RTX 4090 GPU（每GPU批次大小为1），测试阶段使用1张同型号GPU（每GPU批次大小为1）。针对损失函数 $\mathcal{L}=\mathcal{R}(y_i)+\lambda\cdot(\mathcal{D}(x_i,\hat{x}_i)+\beta\cdot\mathcal{P}(x_i,\hat{x}_i))$，分别对平衡系数 $\lambda=1$、8、256、512这4种情况进行训练。其中，当 $\lambda=1$、512时，学习率设为 $l_r=1\times10^{-6}$，当 $\lambda=8$、256时，学习率设为 $l_r=1\times10^{-5}$。在测试阶段，沿用上述λ参数配置，各对比算法也选取4种不同重建质量进行对比。本文采用学习感知图像块相似度LPIPS[26]作为视频帧相似性度量指标，以更贴近人类视觉感知，并通过*R-D*曲线反映所提算法与对比算法的客观性能。

3.2 对比实验结果及分析

采用的对比算法包括基于深度学习的视频编码算法，如条件引导的视频编码算法DCVC[9]、DCVC-TCM[10]、DCVC-HEM[11]、DCVC-DC[12]和DCVC-FM[13]，*R-D*曲线如图7所示，横轴为编码每像素需要的比特率（Bit per pixel, BPP），纵轴为学习感知图像块相似度LPIPS[26]。由图可知，所提算法的LPIPS值在各个数据集中均显著低于对比算法，并且随着BPP的增加，所提算法的LPIPS值下降速度较快。具体地，对于HEVC数据集，相同的码率区间，LPIPS值平均下降18.37%，相同LPIPS值，BPP平均下降11.55%。对于3种空天数据集，相同的码率区间，LPIPS值平均下降36.49%，相同LPIPS值，BPP平均下降31.20%。由此可知，所提算法重建图像的感知质量更接近原始图像，并且在不同比特率下具有较好的适应性。不仅在传统视频编码数据集HEVC[30]上表现良好，在经典的空天数据集（无人机航拍视频UAVDT[27]、ERA[28]、AU-AIR[29]）上，CGDM-VC算法也重建出高质量的视频帧。这充分验证了该算法在保证视频质量的情况下，有效降低了编码比特率，满足了无人机需要大量拍摄与传输视频时对压缩质量以及码率的需求。

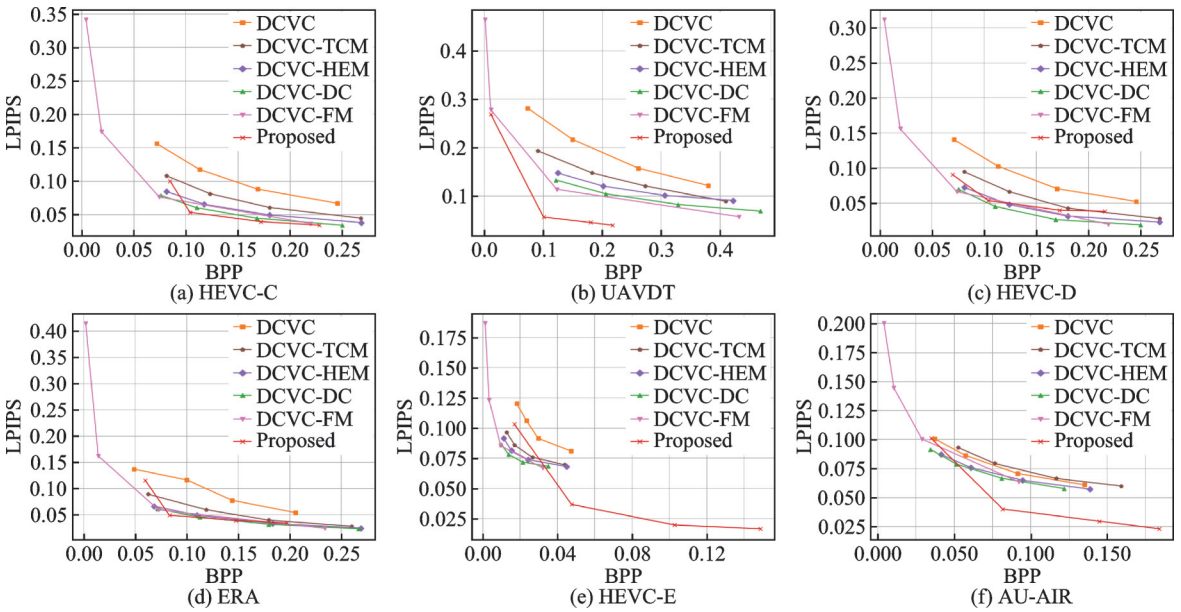

图7 视频编码对比实验客观性能评估结果

Fig.7 Objective performance evaluation results of video compression comparison experiment

所提算法与对比算法在HEVC-C[30]数据集和ERA[28]数据集上的主观对比结果分别如图8和图9所示。由图8可知,对比算法的重建帧中存在大量的模糊伪影,未能清晰重建地板上的纹路,而所提算法在压缩比相当的情况下,成功重建了地板上的钉子,地板的纹路也更加接近原始视频帧。由图9可知,原有的视频算法重建的火灾现场烟雾模糊,而所提算法在更大的压缩比例下,成功重建了烟雾所遮挡的树枝细节,更准确地反映了烟雾本身的浓度和飘动趋势。综上所述,所提算法不仅在传统数据集上具有良好的重建效果,对无人机拍摄的视频帧也具有良好的重建效果,验证了相较于DCVC[9]等经典算法的有效性和多场景应用的可靠性。

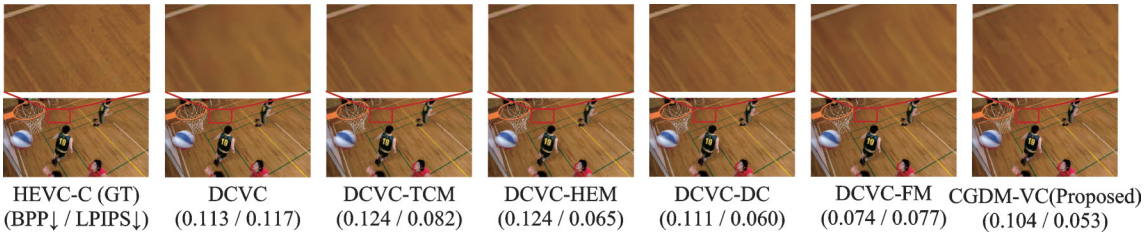

图8 HEVC-C重建质量对比结果

Fig.8 Comparison results of reconstruction quality on HEVC-C

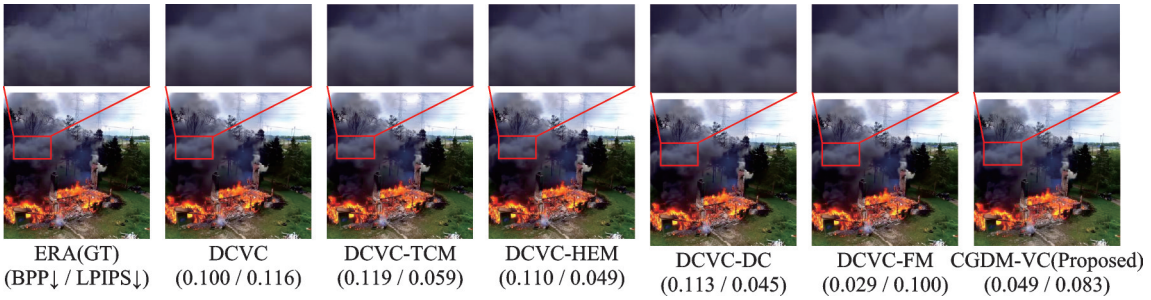

图9 ERA重建质量对比结果

Fig.9 Comparison results of reconstruction quality on ERA

所提算法与对比算法在无人机视频数据集(crossroads和night)上的主观对比结果分别如图10和图11所示。在crossroads场景中,对比算法DCVC等在低码率(<0.1 BPP)下重建的路面白色虚线和水迹区域明显模糊,所提算法白色虚线边缘清晰连续,水迹纹理与原图一致。在night场景中,对比算法DCVC等在低码率(<0.1 BPP)下重建的路面纹理过度平滑,白色虚线边缘弥散,所提算法路面颗粒细节保留完整,白色虚线锐度接近原始帧。由此可知,所提算法在低码率情况下可以保留完整的无人机视频细节,满足了在视频监控、夜间巡检等场景中的需求。

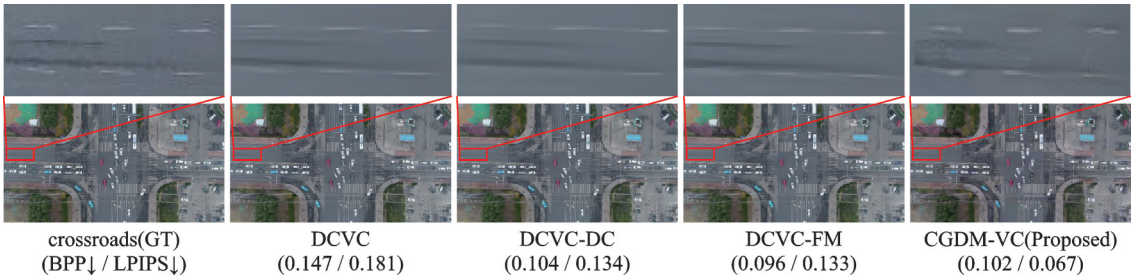

图10 无人机数据集crossroads视频的重建质量对比结果

Fig.10 Comparison results of reconstruction quality on crossroads videos in drone dataset

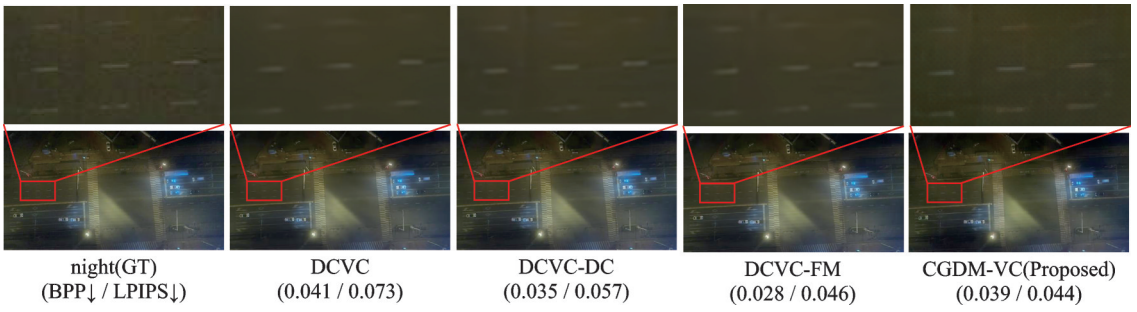

图 11 无人机数据集 night 视频的重建质量对比结果

Fig.11 Comparison results of reconstruction quality on night videos in drone dataset

3.3 消融实验结果及分析

为了验证所提算法中自适应时空重要性引导单元以及隐式帧间对齐模块的性能，本文在 HEVC[30] 的 Class-C、Class-D、Class-E 这 3 个数据集上完成了相应的消融实验，实验结果分别如图 12 和图 13 所示。

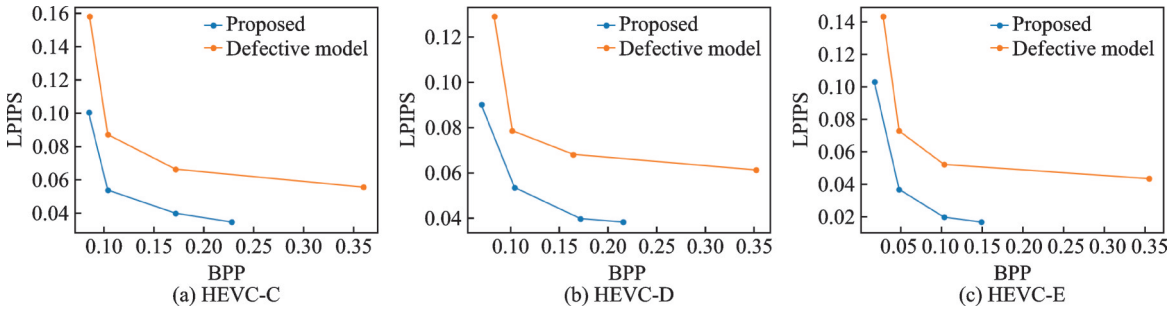

图 13 隐式帧间对齐模块消融实验客观结果

Fig.13 Objective performance results of implicit inter-frame alignment module

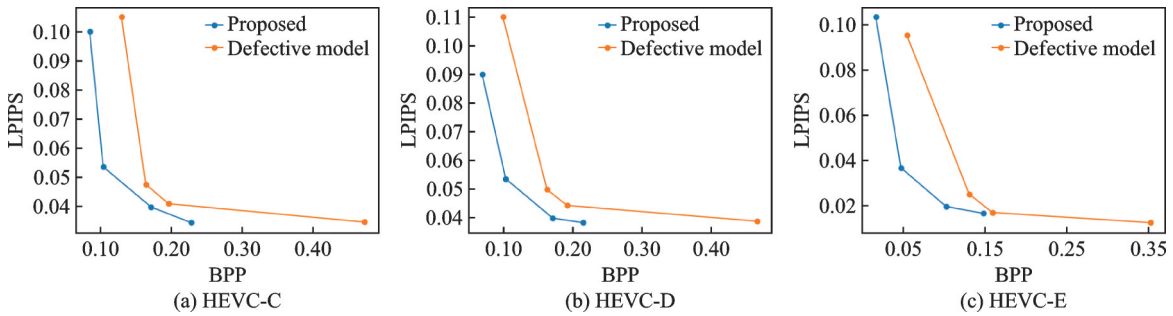

图 12 时空重要性引导单元消融实验客观结果

Fig.12 Objective performance results of spatio-temporal importance guided unit

由图 12 可知，如果缺失帧间时空重要性引导单元提供的引导信息，消融模型在 3 个数据集上的 BPP 明显增加、重建视频的 LPIPS[26] 明显增大。这些结果验证了时空重要性引导单元的有效性。

由图 13 可知，若缺少隐式帧间对齐模块，重建视频帧因缺少帧间参考信息，消融模型在 3 个数据集上的 BPP 与 LPIPS 明显增大。这说明隐式帧间对齐模块在重建高质量视频帧的优势，验证了该模块的有效性。

3.4 计算复杂度分析

对比CGDM-VC与DCVC[9]、DCVC-TCM[10]、DCVC-HEM[11]、DCVC-DC[12]、DCVC-FM[13]的时间与空间的复杂度。在NVIDIA GeForce RTX 4090 GPU、Intel(R) Xeon(R) Gold 6426Y CPU计算平台上,各算法对分辨率为832×448的图像进行编码和解码的平均运行时间与显存空间的占用情况如表1所示。其中,“参数量”表示P帧模型的参数量,“时间”计算了每张图像的平均推理时间,“显存占用”是测试所占用的显存空间,“LPIPS”是$\lambda=512$的情况下重建帧与原视频帧之间感知图像块相似度的平均值。如表1所示,DCVC系列算法呈现出不同的性能表现。DCVC-TCM算法在显存占用方面降至2 244 MiB;DCVC-HEM算法实现了0.24 s的最短运行时间;DCVC-DC算法的参数量为29.562×10^6,LPIPS值为0.082 2,降低了模型的参数量、提升了视频重建质量。

表1 平均运行时间与显存空间占用

Table 1 Average running time and video memory usage

对比算法	参数量/10^6	时间/s	显存占用/MiB	LPIPS↓
DCVC[9]	8.797	0.34	3 670	0.156 9
DCVC-TCM[10]	10.213	0.28	2 244	0.120 5
DCVC-HEM[11]	16.712	0.24	2 264	0.101 0
DCVC-DC[12]	29.562	0.28	5 608	0.082 2
DCVC-FM[13]	25.356	0.31	3 072	0.113 4
CGDM-VC	184.920	0.48	6 154	0.039 4

本文提出的CGDM-VC算法引入了庞大参数量的扩散模型,增加了计算复杂度。然而该算法的推理时间和显存占用与对比算法处于同一数量级,并在LPIPS值上表现最佳,相较于LPIPS表现次优的DCVC-DC算法,其LPIPS性能提升了52.07%。这表明在不显著增加计算资源消耗的前提下,本文算法实现了更复杂更精确的推理能力,适应于无人机视频编码应用场景。

4 结束语

本文提出了一种基于条件引导扩散模型的视频编码算法CGDM-VC,用于解决低码率环境下视频数据的感知质量降低的问题。该算法设计了隐式帧间对齐策略,通过扩散模型有效捕捉了帧间的潜在特征,优化了帧间信息的处理与重建能力。同时,设计了自适应时空重要性编码器,通过更加灵活的压缩方式自适应地调整压缩参数,进一步提高了视频的重建质量。实验结果表明,该算法在通用数据集和无人机拍摄数据集上,在低码率条件下有效地重建了高感知质量的视频帧。与现有的深度学习编码方法相比,该算法在感知质量和压缩效率上均表现出色,具有较强的实用性。然而,扩散模型的计算复杂度相对较高,下一步将继续优化模型结构,考虑对模型进行剪枝和量化来降低计算复杂度,利用硬件加速(如GPU或TPU)提升推理速度,以更好地应对复杂多变的空天地海场景的实时需求。

参考文献:

[1] 李冬青. 空天地海一体化信息网络中高时效传输技术研究[D]. 哈尔滨: 哈尔滨工业大学, 2023.
LI Dongqing. Research on timeliness-sensitive transmission techniques for space-air-ground-sea integrated network[D]. Harbin: Harbin Institute of Technology, 2023.

[2] 王一兆. 面向新一代视频编码标准的码率控制优化算法研究[D]. 北京: 北京邮电大学, 2024.
WANG Yizhao. Research on rate control optimization algorithm for the new generation video coding standard[D]. Beijing: Beijing University of Posts and Telecommunications, 2024.

[3] 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. Review of very low bitrate image compression techniques[J]. Journal of Data Acquisition and Processing, 2025, 40(1): 102-116.

[4] 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.

[5] 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.

[6] 王之琛. 视频编码技术的应用及发展趋势[J]. 中国传媒科技, 2024(7): 134-137,146.
WANG Zhichen. Application and development trends of video compression technology[J]. China Media Technology, 2024(7): 134-137,146.

[7] 朱秀昌,唐贵进. 基于学习的视频编码技术进展[J]. 南京邮电大学学报(自然科学版), 2022, 42(2): 1-12.
ZHU Xiuchang, TANG Guijin. Advances in learning-based video coding technologies[J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition), 2022, 42(2): 1-12.

[8] 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 Bench: IEEE, 2019: 11006-11015.

[9] LI J, LI B, LU Y. Deep contextual video compression[C]//Proceedings of the Neural Information Processing Systems. Virtual: IEEE, 2021: 18114-18125.

[10] SHENG X, LI J, LI B, et al. Temporal context mining for learned video compression[J]. IEEE Transactions on Multimedia, 2023, 25: 7311-7322.

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

[12] 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. Vancouver, Canada: IEEE, 2023: 22616-22626.

[13] 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. Seattle: IEEE, 2024: 26099-26108.

[14] SHENG X, TANG C, LI L, et al. NVC-1B: A large neural video coding model[J/OL]. arxiv.org/pdf/2407.19402, 2024-7-28.

[15] SHENG X, LI L, LIU D, et al. VNVC: A versatile neural video coding framework for efficient human-machine vision[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(7): 4579-4596.

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

[17] CHANG J, ZHAO Z, JIA C, et al. Conceptual compression via deep structure and texture synthesis[J]. IEEE Transactions on Image Processing, 2022, 31: 2809-2823.

[18] MENTZER F, AGUSTSSON E, BALLÉ J, et al. Neural video compression using GANs for detail synthesis and propagation [C]//Proceedings of the European Conference on Computer Vision. Tel Aviv: IEEE, 2022: 562-578.

[19] DUNDAR A, GAO J, TAO A, et al. Progressive learning of 3D reconstruction network from 2D GAN data[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 46(2): 793-804.

[20] LAN C, YAN H, LUO C, et al. GAN-based multi-view video coding with spatio-temporal EPI reconstruction[J]. Signal Processing: Image Communication, 2025, 132: 117242-117250.

[21] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[C]//Proceedings of the Neural Information Processing Systems. Virtual: IEEE, 2020: 6840-6851.

[22] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 10684-10695.

[23] VOLETI V, JOLICOEUR-MARTINEAU A, PAL C. MCVD-masked conditional video diffusion for prediction, generation, and interpolation[C]//Proceedings of the Neural Information Processing Systems. Law Street: IEEE, 2022: 23371-23385.

[24] SONG Y, ERMON S. Improved techniques for training score-based generative models[C]//Proceedings of the Advances in Neural Information Processing Systems. Virtual: IEEE, 2020: 12438-12448.

[25] LUO Y, YANG Z. DynGAN: Solving mode collapse in GANs with dynamic clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(8): 5493-5503.

[26] ZHANG R, ISOLA P, EFROS A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 586-595.

[27] DU D, QI Y, YU H, et al. The unmanned aerial vehicle benchmark: Object detection and tracking[C]//Proceedings of the European Conference on Computer Vision. Munich: IEEE, 2018: 370-386.

[28] MOU L, HUA Y, JIN P, et al. ERA: A data set and deep learning benchmark for event recognition in aerial videos [Software and Data Sets][J]. IEEE Geoscience and Remote Sensing Magazine, 2020, 8(4): 125-133.

[29] BOZCAN I, KAYACAN E. AU-AIR: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance[C]//Proceedings of the IEEE International Conference on Robotics and Automation. Paris: IEEE, 2020: 8504-8510.

[30] BOSSEN F. Common test conditions and software reference configurations[C]//Proceedings of the JCT-VC Meeting. Guangzhou, China: IEEE, 2010: 1-5.

[31] 王崇宇,毛琪,金立标. 基于生成对抗网络的图像视频编码综述[J]. 中国传媒大学学报(自然科学版), 2022, 29(6): 19-28.
WANG Chongyu, MAO Qi, JIN Libiao. Review on image and video coding via generative adversarial networks[J]. Journal of Communication University of China (Natural Science Edition), 2022, 29(6): 19-28.

[32] LI X, LIU Y, CAO S, et al. DiffVSR: Enhancing real-world video super-resolution with diffusion models for advanced visual quality and temporal consistency[J/OL]. https://arxiv.org/pdf/2501.10110, 2025-5-8.

[33] MAO X, JIANG Z, WANG F Y, et al. OSV: ONE step is enough for high-quality image to video generation[J/OL]. https://arxiv.org/pdf/2409.11367?, 2024-9-17.

[34] LI C, JIANG T, WANG H, et al. Optimizing latent variables in integrating transfer and query based attack framework[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 47(1): 161-171.

[35] WANG J, YUE Z, ZHOU S, et al. Exploiting diffusion prior for real-world image super-resolution[J]. International Journal of Computer Vision, 2024, 132(12): 5929-5949.

[36] XUE T, CHEN B, WU J, et al. Video enhancement with task-oriented flow[J]. International Journal of Computer Vision, 2019, 127: 1106-1125.

作者简介:

刘美琴(1980-),**通信作者**,女,教授,博士生导师,研究方向:多媒体信息处理、三维视频处理、视频智能编码,E-mail: mqliu@bjtu.edu.cn。

陈虹宇(2004-),女,本科生,研究方向:视频编码。

周一鸣(2003-),男,本科生,研究方向:视频编码。

倪文昊(2000-),男,博士研究生,研究方向:视频编码。

(编辑:夏道家)