

FFiT:基于FasterViT的快联合帧插值去模糊方法

程之荣, 徐杨

引用本文

程之荣, 徐杨. FFiT:基于FasterViT的快联合帧插值去模糊方法[J]. 计算机科学, 2026, 53(7): 34-44.

CHENG Zhirong, XU Yang. FFiT:Faster Frame Interpolation Transformer Based on FasterViT[J].

Computer Science, 2026, 53(7): 34-44.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[融合动态稀疏性与异构知识蒸馏的Top-k推荐算法](#)

Dynamic Sparsity and Heterogeneous Knowledge Distillation for Top-k Recommendation

计算机科学, 2026, 53(6A): 250700121-9. <https://doi.org/10.11896/jsjcx.250700121>

[融合多尺度特征筛选的轻量露天矿爆堆矿石图像分割算法](#)

Multi-scale Feature Screening-integrated Lightweight Algorithm for Blast Heap Ore Image

Segmentation in Open-pit Mines

计算机科学, 2026, 53(6A): 250500016-8. <https://doi.org/10.11896/jsjcx.250500016>

[基于改进YOLOv8模型的雨雾天气目标识别算法](#)

Rain and Fog Weather Object Detection Algorithm Based on Improved YOLOv8 Model

计算机科学, 2026, 53(6A): 250300090-7. <https://doi.org/10.11896/jsjcx.250300090>

[基于双重抗遗忘机制的轻量化联邦持续学习方法](#)

Lightweight Federated Continual Learning Method Based on Double Anti-forgetting Mechanism

计算机科学, 2026, 53(4): 424-434. <https://doi.org/10.11896/jsjcx.250500116>

[基于改进YOLO算法的学生行为检测方法](#)

Student Behavior Detection Method Based on Improved YOLO Algorithm

计算机科学, 2026, 53(3): 246-256. <https://doi.org/10.11896/jsjcx.241100165>

FFiT: 基于 FasterViT 的快联合帧插值去模糊方法

程之荣 徐 杨

贵州大学大数据与信息工程学院 贵阳 550025

(gs.zrcheng23@gzu.edu.cn)

摘 要 在动态场景图像采集领域,消费级成像设备通常受限于硬件架构缺陷与曝光参数优化不足,长期存在帧率受限与运动模糊共存的复合型退化问题。现有方法在真实模糊场景下存在重建精度不足、模型泛化能力受限,以及计算效率低下等技术瓶颈。对此,提出一种基于 FasterViT 架构的联合帧插值与运动解耦优化框架 FFiT(Faster Frame Interpolation Transformer),旨在实现动态模糊序列的高效时空联合建模。该框架通过构建 CNN+Transformer 混合编码架构,集成了以下关键模块:1)改进型多尺度残差 Transformer 模块(Multi-scale Residual Transformer Block, MRTB),旨在通过时空域自注意力机制解决重建精度问题,并显式构建模糊帧间的事件相关性;2)高质量特征传输网络(High-Quality Feature Transmitter, HQFT),采用跨尺度特征蒸馏机制增强模糊-清晰域转换的语义一致性,以提升模型的泛化能力;3)轻量化动态上采样渲染模块(Dynamic Upsampling and Rendering Module, DYRM),通过可微分动态卷积实现分辨率重建与计算复杂度的解耦优化,解决计算效率低下并实现灵活的分辨率恢复。在 Adobe240 和 RBI 数据集上进行实验,结果表明,FFiT 在解决动态模糊序列的时空联合建模难题方面展现了卓越性能:不仅将模型参数量降低了 70%,还将峰值信噪比(PSNR)较基线模型提升了 2.73 dB,在提升图像质量的同时平衡了计算效率,为解决此类复合退化问题提供了有价值的参考。

关键词: 运动模糊;联合插值去模糊;CNN+Transformer;轻量化

中图分类号 TP751

FFiT: Faster Frame Interpolation Transformer Based on FasterViT

CHENG Zhirong and XU Yang

College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China

Abstract In the realm of dynamic scene image acquisition, consumer-grade imaging devices commonly suffer from a compound degradation problem characterized by coexisting limited frame rates and motion blur. These issues primarily stem from inherent hardware architecture deficiencies and suboptimal exposure parameter optimization. Prevailing methodologies aimed at mitigating these degradations encounter significant technical bottlenecks in real-world blurry scenarios, including insufficient reconstruction accuracy, restricted model generalization capabilities, and low computational efficiency. To address these limitations, this paper introduces FFiT(Faster Frame Interpolation Transformer), a novel joint frame interpolation and motion decoupling optimization framework based on the FasterViT architecture. FFiT is designed to achieve efficient spatio-temporal joint modeling of dynamic blurry sequences. The framework integrates a CNN+Transformer hybrid encoding architecture, incorporating several key modules: 1) An improved multi-scale residual transformer block(MRTB), which leverages spatio-temporal self-attention mechanisms to enhance reconstruction accuracy and explicitly model event correlations between blurry frames; 2) A high-quality feature transmitter(HQFT) module, employing a cross-scale feature distillation mechanism to bolster semantic consistency during the blurry-to-sharp domain conversion, thereby addressing the challenge of model generalization; 3) A lightweight dynamic upsampling and rendering module(DYRM) that utilizes differentiable dynamic convolution to decouple resolution reconstruction from computational complexity, thus tackling computational inefficiency and enabling flexible resolution recovery. Experimental evaluations on the Adobe240 and RBI datasets demonstrate that FFiT exhibits exceptional performance in addressing the intricate spatio-temporal joint modeling of dynamic blurry sequences. Notably, FFiT reduces model parameters by 70% to accommodate computational resource constraints while achieving a 2.73 dB improvement in PSNR(Peak Signal-to-Noise Ratio) compared to baseline models. This research offers a valuable reference for resolving such compound degradation problems by effectively balancing image quality enhancement with computational efficiency.

到稿日期:2025-04-10 返修日期:2025-07-14

基金项目:贵州省科技计划(黔科合成果(2024)重大 004)

This work was supported by the Guizhou Province Science and Technology Plan(Qiankehe Achievements(2024)Major 004).

通信作者:徐杨(xuy@gzu.edu.cn)

Keywords Motion blur, Joint interpolation deblurring, CNN+Transformer, Lightweight

1 引言

在视频采集系统中,快门周期(帧速率的倒数)与曝光时间作为两个核心可控参数,相较于场景内物体运动、光学畸变及传感器噪声等外部干扰变量,其优化配置对视频质量具有决定性影响^[1]。对于消费级相机而言,其帧率调控能力有限,更重要的是,曝光时间往往根据光照条件动态变化且缺乏精确记录,导致实际拍摄的视频帧间曝光差异显著,所产生的运动模糊形态也因此极具多样性和不确定性。这种不可预测的曝光机制对依赖固定或已知曝光假设的去模糊算法构成了重大挑战,因为可变的曝光直接导致了模糊核的复杂化与时变性,使得从模糊观测中精确恢复清晰细节变得尤为困难^[2-3]。

学术界近年来致力于系统研究如何从上述动态变化中提取清晰图像特征,并在该领域取得了突破性进展。自 Jin 等^[3]进行开创性研究以来,研究范式逐步向更具挑战性的方向转移,即通过单张运动模糊图像重建高帧率清晰视频序列。该技术路径可形式化定义为联合去模糊与时域超分辨率^[4],或更精确地表述为模糊时间超分辨率^[5-6]。此类融合算法的应用场景涵盖视频感知质量增强、高帧率慢动作重构^[5],以及高速运动目标解析^[6]等关键领域。

既有研究^[4,7-8]证明了联合学习的优势。DeMFI 模型^[5]通过特征流驱动的自适应形变实现了任意时刻的清晰帧合成,展现了时间插值的灵活性。然而,其依赖于对曝光参数的隐式假设,当实际曝光条件与训练设定存在较大偏差时,模型性能的鲁棒性和泛化能力可能面临严峻考验,难以有效处理

真实世界中常见的复杂模糊模式。VIDUE^[9]等方法通过探索时序关联,构建了事件驱动的端到端框架,并在一定程度上通过学习曝光感知表征展现了对未知曝光时间的鲁棒性。尽管如此,如何在复杂动态场景下,进一步提升模型对此类曝光不确定性的自适应能力,同时减小由动态范围压缩导致的高频细节不可逆损失,实现在有限计算资源下重建高质量、高帧率的清晰视频,仍是该领域亟待解决的问题。

当前工作^[1,9]中,视频去模糊与插帧任务中存在局部细节退化与全局运动失配问题,亟需构建兼具细粒度特征感知与长程时序建模能力的混合架构。传统的卷积神经网络(Convolutional Neural Networks, CNN)在局部特征提取方面具有显著优势,然而仅依赖局部卷积操作往往难以捕捉跨帧或长距离依赖关系。同时,随着视频数据规模与分辨率的不断提升, CNN 在处理时间序列数据时也会面临并行化和全局信息捕捉有限等挑战。而 Transformer 架构依靠自注意力机制,能够高效挖掘序列数据的全局依赖,并提供并行计算能力。在此背景下,本文通过新的方法将 CNN 与 Transformer 架构进行结合,以充分发挥两者在局部特征学习与全局信息建模上的互补优势,从而在运动图像去模糊的上游任务中取得卓越效果。

在合成数据集与真实数据集上进行实验,结果表明,无论从视觉效果还是定量指标(如 PSNR/SSIM)来看,本文提出的 FFiT 模型在保持轻量化的同时,其性能与最新的先进模型相当甚至更优。如图 1 所示,FFiT 可以通过在少量模糊帧之间插入次清晰帧并最终恢复清晰帧的方式,持续生成高质量的去模糊帧。



图 1 FFiT 实现的 30 帧任意模糊插值
Fig. 1 Generating 30 sharp frames from blurred image using FFiT

实验结果表明,本文模型在复杂运动场景下保持了较好的泛化能力和高效性。针对上述研究目标,本文的主要贡献体现在以下 3 个方面。

1) 由于运动模糊常呈现出多尺度、多类型的特征,本文改

进了分层注意力机制,引入了多尺度残差注意力,使模型能够在不同尺度上进行特征提取与融合。通过在各级尺度下自动学习帧间运动及模糊信息,模型不仅可以准确捕捉高速变化场景下的细微纹理,还可在插帧过程中保证运动的连贯性和

完整性。该机制既拓展了模型的感受野,又通过残差累加方式使特征信息的回传更具鲁棒性。

2)为了提升 Transformer 对高质量特征的建模能力,本文提出了 HQFT 模块。在将原始图像输入图像复原子模块之前,HQFT 先通过数据增广、非线性映射等方式对关键特征进行提取与重构,从而有效保留与运动模糊去除密切相关的隐式信息。该过程能够保证 Transformer 处理的 token 具备更丰富、准确的表征,使得模型在时间维度上共享潜在特性,并实现基于时间序列的精准重建。

3)针对不同类型及不同程度的运动模糊,本文提出了基于点采样的上采样方法,并进行了分组采样设计,最终形成了 DYRM 模块。它能够根据图像局部区域的模糊程度和运动形变,自适应地进行动态上采样。该过程不仅能更好地呈现非线性隐式特征,也有助于在重建后期保留更多细节,从而在多样化、复杂化的实际场景中保持对图像时空信息的精细复原能力。

2 相关工作

2.1 图像去模糊

在图像去模糊领域,传统方法大多基于最大后验(Maximum A Posteriori, MAP)优化框架,其核心依赖于对自然图像统计特性的先验建模,包括但不限于总变差正则化、马尔可夫随机场平滑先验、归一化稀疏性约束以及色彩一致性先验等。随着大数据技术的演进,研究范式已从传统先验驱动方法转向数据驱动的端到端学习策略,即利用大规模模糊-清晰图像对训练深度网络以直接建立去模糊映射关系^[10]。这一范式革新不仅推动了多阶段渐进式估计理论的完善^[11],更催生了一系列新型网络架构,典型代表包括 Chi 团队提出的多尺度 U-Net 模型,以及 Ren 等提出的基于深度先验的无监督去模糊框架^[12]。

针对视频去模糊任务,学术界通过引入递归计算机制^[13]、光流估计技术^[14-15]以及可变形卷积算子^[16],有效实现了时序运动补偿的数学建模,显著提升了动态场景的复原质量。当前主流方法普遍存在运动信息重复计算的问题,因此融合视频帧插值与去模糊操作的联合优化策略逐渐成为研究焦点,其核心目标在于降低计算复杂度并提升算法执行效率。

从技术发展脉络分析,图像与视频去模糊方法经历了从基于模糊核估计的传统模型^[16]到深度学习驱动的直接复原方法^[16]的范式转变。在此进程中,研究者在网络架构设计层面进行了系统性探索:一方面通过卷积神经网络(CNN)、递归神经网络(RNN)及生成对抗网络(GAN)等基础架构构建多样化解决方案;另一方面提出多尺度特征融合与多阶段渐进式处理策略,以满足不同应用场景需求。多尺度 CNN 架构在静态图像去模糊任务中具有显著性能优势。例如,Zhao 等^[17]提出的基于残差注意力和多特征融合的去模糊网络,通过双分支网络结构有效补偿信息损失并复原图像细节;而 Zamir 等^[18]则通过多级渐进式框架验证了层次化特征重构的有效性。此外,针对视频序列的时序关联特性,Wang 等^[16]与 Pan 等^[14]提出的可变形卷积-光流联合估计模型,实现了运动补偿精度与计算效率的协同优化。

Transformer 架构在低级视觉任务中的突破性进展为去模糊领域注入了新的研究动力。以 Restormer^[19],RVRT^[20]及 VDTR^[21]为代表的先进模型,不仅在网络拓扑设计与训练范式上取得重要突破,更为后续算法创新提供了多维度的技术启示与理论支撑。

2.2 CNN + Transformer

Transformer 架构在自然语言处理领域的突破性应用,为计算机视觉任务开辟了新的技术路径,这一范式迁移被公认为计算机视觉发展史上的关键转折点。作为典型代表,Vision Transformer(ViT)通过引入图像块级注意力机制,有效验证了全局建模在视觉理解任务中的理论可行性。其架构与学习机制的简洁性,使其迅速成为视觉表征学习领域的研究焦点。针对 ViT 的性能优化,学术界主要沿着架构创新方向展开系统性研究:DeiT^[22]采用卷积与自注意力的混合架构强化模型归纳偏置;LeViT^[23]基于轻量化设计原则重构网络参数分布;多级金字塔结构^[24]的引入实现了 CNN 与 Transformer 的深度特征融合;T2T-ViT^[25]通过令牌重组策略优化信息传递路径;GCViT^[26]则构建了全局上下文感知的注意力计算范式。相关衍生研究进一步拓展了网络层次设计与注意力机制的理论边界^[27-28]。

随着模型规模与训练数据量的指数级增长,注意力机制固有的二次计算复杂度导致 Transformer 架构面临严峻的计算效率瓶颈^[29]。针对此问题,VisFormer^[30]提出将计算密集型 Transformer 架构重构为等效卷积算子,显著提升了视觉识别任务的推理速度;CrossViT^[31]通过多尺度特征融合策略,结合不同粒度的图像块划分与线性复杂度注意力机制,成功将时间复杂度与内存占用量降低至可接受范围。然而,现有方法在保持模型表征能力与降低计算代价之间仍需权衡,如何构建理论最优的注意力计算范式仍是当前研究的重点攻关方向。

在近期技术突破中,NVIDIA 研究团队提出的 FasterViT 混合架构,针对计算机视觉场景的高吞吐量图像处理需求进行了系统性优化。该模型创新性地融合了卷积神经网络在高效局部特征提取方面的优势,以及视觉 Transformer 在全局上下文建模中的核心能力,构建了分层注意力(Hierarchical Attention, HAT)机制。通过将全局自注意力分解为多级子模块,该方法在保持模型表征能力的同时,将计算复杂度降至近似线性水平。该团队提出的混合架构通过参数共享与结构重参数化技术,实现了模型参数量与分类精度的有效权衡,为工业级视觉系统的部署提供了新的技术范式。

2.3 模糊插帧

在视频处理领域,提升时间分辨率(即帧率上转换)的技术研究受到学者的持续性关注,其对于视频压缩等应用场景具有重要工程价值。随着深度学习技术的快速发展,基于数据驱动的视频帧插值方法逐渐成为计算机视觉领域的研究热点,特别是在中心帧插值与多帧插值方向取得突破性进展。例如,Liu 等^[32]针对抖动视频的前景目标提取问题,提出了结合稳像、补帧以及改进 VIBE 算法的策略,通过稠密光流法生成中间帧,有效提升了动态场景下目标提取的完整性; Bao 等^[33]引入深度估计与双向光流建模来解决遮挡区域的插值

难题;Lee 等^[34]则通过核方法与光流法的协同架构,结合对抗性损失评估插值帧的视觉自然度。相较之下,多帧插值任务需构建更为复杂的运动轨迹模型,如 Xu 等^[35]提出的二次插值策略,以及 Chi 等^[36]基于立方体运动模型实现的复杂场景高精度插值。

在处理含运动模糊的视频数据时,基于光流或显式运动估计的传统方法存在显著局限性,原因在于模糊效应会直接导致运动表征失准。这促使研究者转向模糊插帧技术,其核心在于挖掘运动模糊过程中隐含的时空信息来重建真实运动序列。Jin 团队^[3]首次提出通过单帧模糊图像重建清晰视频序列的可行性方案,但受限与方向模糊的不可逆性,重建精度不足。后续工作中,Jin 等^[3]创新性地引入成对阶不变损失函数以减小方向模糊造成的影响;Purohit 等^[37]采用自监督策略学习隐式运动表征,而 Argaw 团队^[38]则通过多分支空间变换网络处理视频模糊场景。针对模糊视频插值任务,现有方法呈现技术路线分化:Jin 等^[3]采用去模糊-插值级联架构,Shen 等^[4]设计金字塔递归网络抑制误差累积效应,Argaw 等^[38]则维持光流估计框架实现端到端插值。Oh 团队^[5]提出的 DeMFI 框架通过流引导注意力特征与递归增强机制的协同作用,实现了低帧率模糊视频到高帧率清晰视频的高效转换。

最新研究趋势表明,联合优化帧插值与去模糊任务的集成式方法成为重点突破方向。Shen 等^[4]开发的金字塔网络架构可同时支持中心帧插值与去模糊任务,Argaw 等^[38]则通过光流引导注意力与外推机制的融合以提升重建质量。需指出的是,现有方法普遍基于固定曝光时间的理论假设,而实际应用场景中曝光参数会随拍摄环境与设备特性动态变化,这使得现有方法的工程适配性仍存在空白,亟待突破。本文针对现有方法在真实模糊场景下重建精度不足、模型泛化能力受限及计算效率低下的技术瓶颈进行了一些尝试。

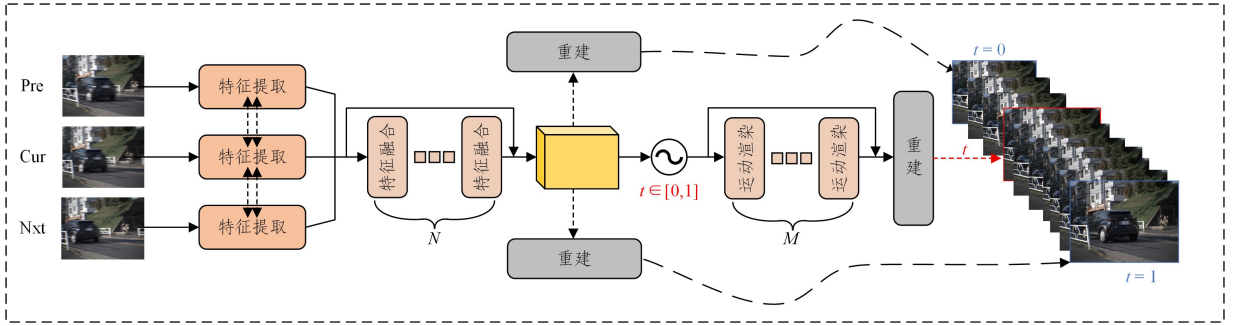
3 本文模型

假设给定的视频序列包含 T 帧,记作 $y = \{y_t\}_{t=1}^T$ 。视频捕捉过程中,单次快门周期 Δt 由曝光阶段 Δt_e 和有效读出阶段 Δt_r 两部分组成,即 $\Delta t = \Delta t_e + \Delta t_r$ 。在第 t 帧, y_t 可以视为每个瞬间 τ 的潜在图像 x_τ 在曝光阶段 $[t \cdot \Delta t, t \cdot \Delta t + \Delta t_e]$ 的积分,记为:

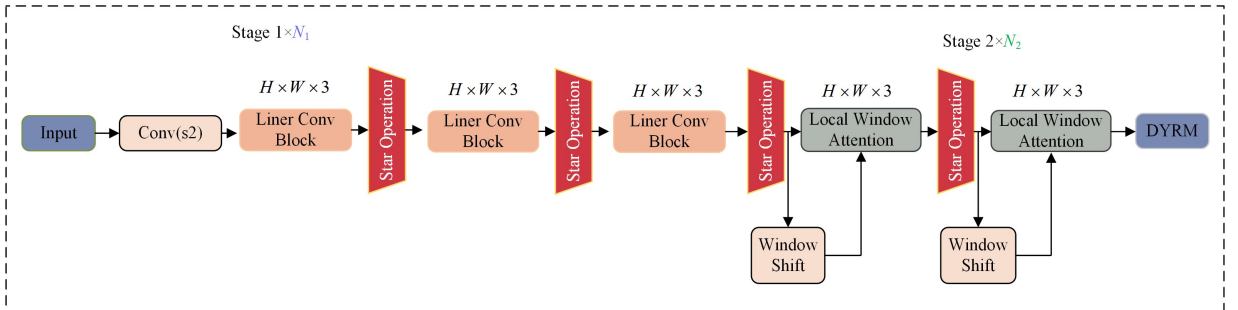
$$y_t = \frac{1}{\Delta t_e} \int_{\tau=t \cdot \Delta t}^{t \cdot \Delta t + \Delta t_e} x_\tau d\tau \quad (1)$$

在消费级相机中,若快门时间较长,则曝光时间在 Δt 中占比较大,从而导致动态场景中的视频往往帧率偏低且伴随着明显的动态模糊。联合视频插值去模糊的目标是在观测到的视频 y 中恢复包含清晰帧的高帧率序列 $x = \{x_t\}_{t=1}^T$ 。此时需要考虑快门周期 Δt 已知但曝光时间 Δt_e 未知,而 Δt_e 与运动模糊强度直接相关。与 DeMFI^[5]采用连续帧直接平均合成模糊(忽略读出阶段)的做法相比,式(1)更为贴近实际拍摄过程,在曝光时间内对局部场景进行了更真实的积分。这种策略可以在插帧过程中充分利用潜在的运动信息,并在合成的模糊帧中补足相邻时刻的精细变化,为后续的锐化提供额外约束。

图 2(a)所示的 FFiT 框架基于 CNN 与 Transformer 相结合的思路,旨在针对给定曝光时间区间,通过插值手段从模糊视频中恢复运动清晰、帧率更高的序列。参考既有研究中的 FasterViT^[39],本文在兼顾模型性能与较高吞吐量的思路下,采用了多尺度残差注意力机制,并进一步在网络中嵌入对相邻帧插值的学习过程。FasterViT 中的分层注意力模块能够有效捕捉局部区域间的交叉窗口交互,同时在全局范围内模拟长距离的空间依赖性。在此基础上,该模块既可提取图像的低级边缘信息,又可捕捉视频中较高级的语义信息。



(a) 模型框架



(b) HQFT 框架

图 2 模型框架和 HQFT 框架

Fig. 2 Model framework and HQFT framework

为适配视频处理场景的高通量特性,本文在 FasterViT 的主干上引入残差结构,并将其分为共享时间特征提取阶段 F_N 与任意运动渲染阶段 F_M 两部分。 F_N 内部包含 N 个多尺度残差注意力模块(MRTB),中间穿插若干特征传递子模块; F_M 与 F_N 的模块堆叠方式相似,但功能定位不同。残差结构的输出再经由额外的卷积层处理,以保持与初始输入相同的维度。在这两部分的残差结构外,本文还提出了两个独立模块:1)用于轻量化特征提取与数据增广的 HQFT,该模块能够在图像进入后续恢复模块之前,先完成特征映射与时间信息融合;2)DYRB,其通过动态上采样有效应对不同种类及不同程度的运动模糊,实现对复杂运动细节的精确重建。此外,为了表示模型整体的处理流程,将 FFiT 的输出形式记作:

$$x = F_R(F_M(F_N(x_t))) \quad (2)$$

其中, F_R 表示最终的恢复与融合操作,用于输出复原后的视频帧。整个网络的训练过程主要采用 L_1 损失来监督重建结果,具体形式如下:

$$L = L_1(\hat{x}_s', x_s') + \lambda(L_1(\hat{x}_s^0, x_s^0) + L_1(\hat{x}_s^1, x_s^1)) \quad (3)$$

其中, $\hat{x}_s'$ 表示由模型生成的修复图像, x_s' 为相应数据集中的清晰图像, λ 为权衡插值与去模糊约束的超参数。在训练阶段,时刻 t 仅从可用的离散时间节点随机采样,测试时可在 $[0, 1]$ 范围内自由取值。通过这种损失设计,模型能够在不同时间位置上学习生成更具时序连贯性的锐利视频帧,同时兼顾去模糊与插帧的精度。

与现存方法相比,FFiT 在充分利用 CNN 提取细节的优势与 Transformer 全局建模能力的同时,分别在特征增广(HQFT)与动态上采样(DYRB)方面进行了针对性设计,以更好地捕捉和重建复杂运动场景。

3.1 MRTB 模块

在图像去模糊与插帧的联合任务中,高效且准确地提取多尺度特征对于重建复杂运动场景而言至关重要。受到 Zhong 等^[1]的启发,本文基于分层注意力的理论框架,设计并实现了多尺度残差注意力模块 MRTB。如图 3 所示,该模块在网络主干中通过残差结构衔接多尺度特征提取,并采用卷积层对多尺度特征进行融合,从而输出与初始输入具有一致形状的特征表示。为适配多尺度处理场景的异构性需求,本文首先对形状为 $R^{C \times H \times W}$ 的原始输入特征进行尺度插值,使其在尺度级别 $s \in \{1, \dots, S\}$ 下分别映射至 $R^{C \times H/r^{s-1} \times W/r^{s-1}}$ 的空间分辨率,其中 r 表示缩放比例。通过在多个尺度上进行并行或串行的注意力运算,网络可以更充分地捕捉不同尺度下的运动模糊。设 S 为 3, r 为 2,意味着可通过多步缩放来提取由粗到细的时空信息。为定量评价多尺度策略所带来的计算代价,记 Trans 表示单尺度 Transformer 的计算量,则 MRTB 的整体复杂度近似为:

$$\Omega(\text{MRTB}) \approx \frac{1 - (1/r)^{2S}}{1 - (1/r)^2} \Omega(\text{Trans}) \quad (4)$$

由式(4)可知,假设 $S=3, r=2$,那么额外的计算量显著低于原总计算量的 $1/3$ 。多尺度处理不仅有助于分离来自不同深度层的运动模糊特征,也能在无需显式估计运动轨迹的前提下,充分融合相邻帧的时间特征并提升对快速或剧烈运动场景的适应性。

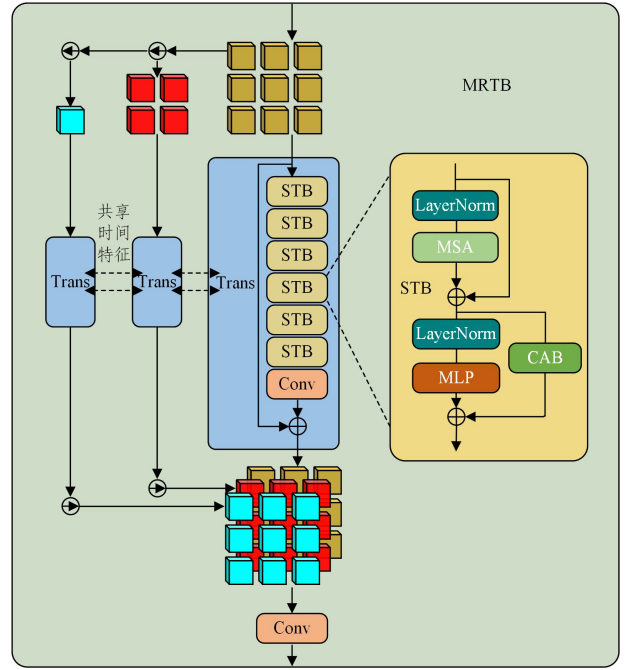


图 3 MRTB 示意图

Fig. 3 Diagram of MRTB

在确保模型高效率运行的同时,MRTB 架构通过残差结构与多头注意力机制的层级化融合,构建了局部特征提取与全局上下文感知的双重优化路径,进而形成多尺度信息的动态表征体系。分层注意力将原本复杂的全局自注意力分解为计算成本更低的多层注意力,借助窗口注意力提高运算并行度及整体吞吐量。残差模块有助于聚合来自相邻帧的插值特征,避免运动过程中的时序或语义扭曲。而在图像修复与重建中,多尺度提取到的低级边缘信息与高级语义信息缺一不可,两者共同决定了输出帧的视觉质量与还原度。

在实验验证中发现,基于分层注意力机制的模型架构在图像重建指标上表现欠佳。具体而言,如图 4(a)所示,尽管该机制通过窗口注意力提升了局部特征提取效率,但其设计存在关键参数复用问题:每个窗口内的注意力值既承担局部结构建模任务,又负责全局表征构建。这种双重功能分配导致跨窗口特征在传递过程中出现表征信息衰减现象,尤其在处理高频纹理细节时,特征流的连续性难以得到有效保障。在图像分类与目标识别任务中,得益于全局-局部协同注意机制的优势,该架构展现出良好的性能表现。然而,图像重建任务对特征保持能力提出了更高要求,需要精确捕捉亚像素级纹理变化及隐式运动轨迹。

基于上述观察,本文经多次实验验证与权衡后,将原分层注意力模块替换为滑动窗口注意力机制。如图 4(b)所示,滑动窗口注意力通过在相邻窗口间有序移动并累积上下文信息,不仅能够更完整地保留细节特征,还可提升全局建模的稳定性。为进一步强化对通道间信息的整合,模型在滑动窗口注意力后引入了通道重建单元(Channel Restore Block, CRB),通过对通道维度进行重整与补偿,整体提升了去模糊与插帧的精度。需要指出的是,尽管这种设计略微增加了模型参数量,但从实验结果来看,其带来的重建质量提升十分显著,更符合图像修复与重建对高保真度和边缘细节的需求。

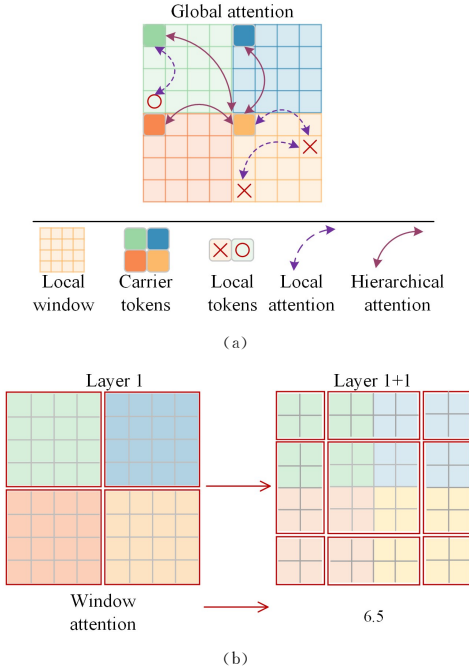


图4 分层注意力与滑动窗口注意力

Fig. 4 Hierarchical attention and shift window attention

MRTB模块通过多尺度残差注意力机制实现了模糊场景的多分辨率表征学习,在引入滑动窗口注意力与CRB架构后,显著提升了图像重建的鲁棒性与算法收敛效率。基于实验数据与可视化分析,后续章节将从量化指标(PSNR/SSIM)与主观感知两个维度,系统论证该模块在视频时序去模糊与帧率上转换复合任务中的技术优越性。

3.2 HQFT 模块

在传统的CNN与Transformer混合框架中,卷积模块与注意力模块往往分别侧重于不同频段的信息提取;卷积模块更倾向于保留高频细节,而注意力模块更善于捕捉低频全局特征。

通常的做法是将这两类模块以级联的方式相结合,再通过简单的下采样操作来衔接。尽管此类方法能够在一定程度上融合两种计算模式,但仍会存在特征对接不稳定、计算资源浪费以及部分特征丢失等问题。为充分利用CNN与Transformer的互补优势,本文在FFiT框架的残差注意力机制(基于窗口注意力)基础上,结合深度可分离卷积的核心思路^[40],提出了高质量特征传输器HQFT,以更有效地在低维空间中生成兼具高频与低频信息的关键特征(即Token),并将其传递给后续的注意力模块。

HQFT模块包括两个阶段。如图2(b)所示,Stage 1旨在生成高质量Token;Stage 2则负责在特征融合过程中保持特征的充分表征,以保证后续的注意力计算能够整合多尺度且多频段的信息。

具体而言,Stage 1在图像进行简单下采样之后,通过图5中展示的基于深度可分离卷积的线性卷积算子对输入特征进行重构与数据增广。此过程可表示为:

$$\hat{x} = \text{GELU}(\text{BN}(\text{LinearConv}_{\text{stage1}}(x))) \quad (5)$$

其中, $\text{LinearConv}_{\text{stage1}}$ 代表基于深度可分离网络思想的线性算子。该线性操作不仅有助于突出图像中的重要特征,使之在

整体特征图中占据更显著的位置,而且在计算开销上远低于传统卷积,其复杂度近似为:

$$\alpha \times C_{\text{out}} \times H \times W \quad (6)$$

而传统深度可分离卷积则需要:

$$(K^2 + C_{\text{out}})(C_{\text{in}} \times H \times W) \quad (7)$$

与之相比,线性操作在保持或增强关键特征的同时,显著降低了冗余运算量。该设计一方面为后续的注意力模块保留了更丰富的边缘与语义信息,另一方面减少了在多尺度卷积层叠加时对计算资源的消耗。Stage 2在初步特征增广后, $\text{StarConv}_{\text{stage2}}$ 将经过BN(Batch Normalization)的输出,并与初始输入进行残差式融合,表达式如下:

$$x = \text{BN}(\text{StarConv}_{\text{stage2}}(\hat{x})) + x \quad (8)$$

如图5所示, $\text{StarConv}_{\text{stage2}}$ 同样基于深度可分离卷积的核心理念,并在处理过程中将每个元素视作一个独立Token,结合通道重建单元CRB实施成对特征乘法核函数,从而在低维空间中生成高维特征,实现非线性隐式映射。如算法1所示,StarOperation可以通过多次星型计算,实现特征空间维度呈指数式扩大,明显增强了对细节纹理和非线性运动的捕捉能力。这种方法在单层计算中并不会显著增加运算量,却能在多层网络堆叠时构造出多重隐性维度,从而实现更丰富的特征表达。

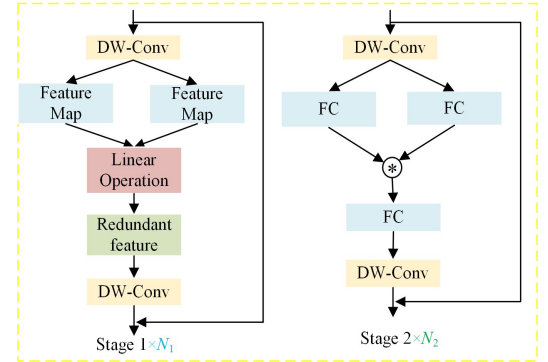


图5 线性卷积与星型卷积

Fig. 5 Linear convolution and star convolution

算法1 StarOperation 的特征映射能力

1. 单层网络的星型计算

$$w_1^T x * w_2^T x = \left(\sum_{i=1}^{d+1} w_1^i x^i \right) * \left(\sum_{j=1}^{d+1} w_2^j x^j \right) = \sum_{i=1}^{d+1} \sum_{j=1}^{d+1} w_1^i w_2^j x^i x^j$$

2. 经过星型计算可以得到多个多项式组合:

$$\frac{(d+2)(d+1)}{2} \approx \left(\frac{d}{\sqrt{2}} \right)^2, d \gg 2$$

且每个多项式都与 x 呈现非线性关系,表明组合之间是单独的隐性维度。

3. 经过 n 层网络可以隐性获得特征空间:

$$O_n = W_{n,1}^T O_{n-1} * W_{n,2}^T O_{n-1} \in \mathbb{R}^{\left(\frac{d}{\sqrt{2}}\right)^{2^n}}$$

算法1中指出,若单层网络采用星型计算方式,则可产生 $O(d^2)$ 量级的多项式组合,其中 d 为输入维度;当网络堆叠至 n 层时,最终隐式特征空间维度可达到 $\left(\frac{d}{\sqrt{2}}\right)^{2^n}$ 量级。正是这种对多项式组合的充分挖掘,使得HQFT能够在保持网络轻量化的同时,大幅提升模型的非线性拟合能力。

相较于传统卷积神经网络与Transformer的串联式混合

架构, HQFT 在特征生成与特征融合环节都更具针对性: Stage 1 以线性增广的方式获得高质量 Token; Stage 2 通过星型计算与 CRB 的深度可分离卷积操作进一步增强非线性映射能力, 确保特征在后续的注意力计算中得到充分利用。通过以上改进, HQFT 不仅大幅降低了网络参数量, 也在实际训练中表现出了更强的鲁棒性与稳定性, 为 FFiT 整体框架的高效视频去模糊和插帧打下坚实基础。

3.3 DYRM 模块

在视频处理领域, 上采样技术是提升视频帧率与分辨率的重要手段, 尤其在运动图像去模糊的场景中, 高质量上采样能够显著恢复因运动模糊而丢失的高频细节。传统方法通常采用最近邻或双线性插值等固定规则来进行上采样, 易于实现但难以适应多变的运动模式与拍摄环境。为获得更灵活的效果, 一些研究在特定任务中引入了可学习的插值方式。然而, 这些可学习方法往往存在伪影严重或难以保证图像边界与重要特征稳定保留等问题。近年来, 随着动态网络的兴起, 部分动态上采样器在多种视觉任务中展现出巨大潜力。

同时, 现代网络架构普遍融合多尺度特征, 使得高分辨率特征并不必直接作为上采样过程中的唯一输入, 反而可以将多尺度信息按需组合, 以获得更好的重建效果。

本文提出的 DYRM 模块通过分组点采样和坐标偏移的方式灵活地完成了上采样与重建。如图 6 所示, 与以往依赖动态卷积的大规模计算不同, DYRM 选择了更轻量化的分组点采样思路: 首先对输入特征图进行初步的上采样, 并将通道维度重新排列到空间维度以提升分辨率。随后在 1×1 卷积层中计算各采样点的水平与垂直方向偏移量 $offset$, 该偏移量依据输入特征图自适应生成(见式(9)), 从而为网络在更高分辨率下提供更细粒度的调整空间。将起始坐标 $ground$ 与缩放后的偏移 $offset$ 相叠加, 得到最终采样点集合 $Sample_set$ (见式(10)), 再通过 $grid_sample$ 操作完成对输入特征的插值(见式(11))。在处理过程中, 针对异常坐标采用线性插值, 不仅能确保全局结构的连续性, 还能对局部细节进行灵活修复与增强。该方法在 PyTorch 框架中仅需调用自带的函数即可实现, 大幅降低了额外的计算资源需求。

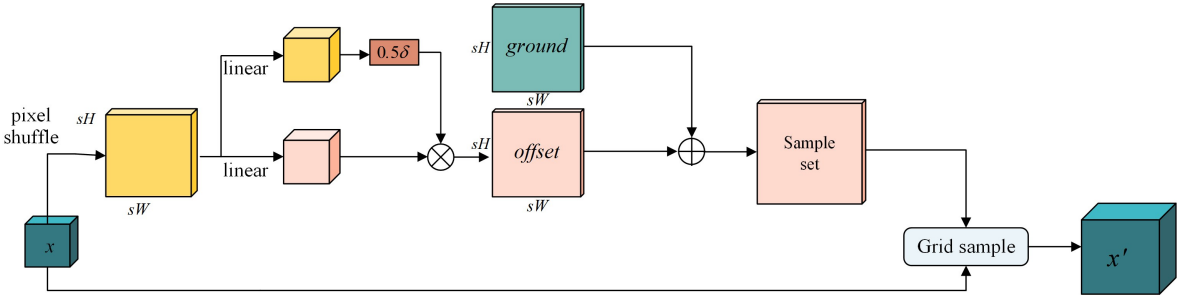


图 6 DYRM 示意图
Fig. 6 Diagram of DYRM

DYRM 的偏移量 $offset$ 由式(9)给出:

$$offset = 0.5 \text{sigmoid}(\text{linear}_1(x)) \cdot \text{linear}_2(x) \tag{9}$$

其中, sigmoid 与线性操作线性组合了输入特征中有价值的纹理或运动信息, 通过 0.5 因子控制更新幅度, 使得网络既能在分辨率放大后进行微调, 也不会过度偏移而导致伪影。随后将该 $offset$ 与初始坐标 $ground$ 相叠加:

上进行了定量和定性的实验。

表 1 数据集介绍

Table 1 Dataset introduction				
数据集	训练集	测试集	图片大小	Epoch
Adobe240	112	8	640×352	800
RBI	50	5	640×480	1600

$$Sample_set = ground + offset \tag{10}$$

再利用常规的 $grid_sample$ 进行插值操作:

$$x' = \text{grid_sample}(x, Sample_set) \tag{11}$$

在此过程中, 每个采样点都可根据实际输入特征自适应地确定插值位置, 从而给网络提供更大的自由度去校正、还原模糊区域的高频信息。

RBI 数据集是专为真实世界的模糊插值方法而设计的, 包含真实世界的模糊与清晰视频对, 由混合摄像系统采集。该数据集包括 55 个混合摄像系统录制的视频, 模糊视频帧率为 25 帧, 清晰视频帧率为 500 帧。模糊图像的曝光时间为 9518 ms, 而清晰图像的曝光时间接近 2 ms。这意味着有 9 个清晰帧对应于一个模糊帧, 两个相邻的模糊帧的死区时间中相差 11 个清晰帧。其数据集的图像分辨率为 640×480 。

综上, DYRM 兼顾灵活性与计算效率, 在不显著增加模型复杂度的前提下实现了对动态场景的高质量重建。得益于分组点采样和偏移自适应生成的设计, DYRM 能够有效应对不同程度的运动模糊与分辨率需求, 为后续的特征融合与视频重建提供了更丰富、准确的细节支撑。结合本文在其他模块(如 HQFT)中的改进, DYRM 进一步推动了 FFiT 整体框架在视频去模糊与帧插值任务上的性能提升。

Adobe240 数据集是一个高帧率视频数据集, 广泛用于视频插帧和时空超分辨率等研究任务。该数据集包括 120 个手持录制的视频, 每个视频的帧速率为 240 FPS, 分辨率为 720×1280 。选取其中的 112 个视频作为训练数据, 其余用于验证和测试。调整分辨率为 640×352 , 并将视频帧组合生成训练序列。

4 实验结果与分析

4.1 数据集介绍

为建立用于视频插值和去模糊的定量性能比较的模拟数据集, 实验通过对高帧率视频进行下采样合成低帧率视频, 设置采样频率 $S=8$, 从而将原始 240 FPS 视频降级为 30 FPS。

如表 1 所列, 本文在 RBI 数据集^[1]和 Adobe 数据集^[41]

4.2 实施细节

本文在 PyTorch 框架中使用 AdamW 优化损失, λ 根据经验设置为 0.5。在初始训练或微调阶段,余弦调度器将学习率从 1×10^{-4} 调度到 1×10^{-6} 。常见的数据增强操作包括翻转、旋转和尺寸为 256×256 的裁剪。针对 Adobe240,采用与以往方法相同的操作,使用 112 个视频用于训练,8 个视频用于测试;RBI 中的 50 个视频用于训练,其余 5 个视频用于测试。在 Adobe240 上训练 BiT 时,批量大小设为 32,持续 800 个 epoch,并在一块 3080 显卡上对其进行微调,再进行 400 个 epoch。由于 RBI 的大小比 Adobe240 小,因此将 epoch 数量加倍,并将批次减少到 8,以获得更多的训练迭代次数。对于其他模型,本文在每个数据集上对其重新训练以进行公平比较。

将 $M=3$ 和 $N=6$ 分别设置为 F_M 和 F_N 的默认 FFiT 模块数量。将自注意力的分类头数量和通道大小分别设置为 6 和 96。

4.3 定量与定性分析

本文将 FFiT 与先前先进的级联方法和光流联合方法进行了比较。对比方法包括:去模糊模型 EDVR^[16]和插值模型 XVFI^[17]组成的级联方法,以及 Jin 等提出的方法^[3]、RPF4^[4]、DeMFI^[5]联合去模糊方法。通过与多个最新的模型进行对比,验证了本文模型的先进性。

4.3.1 在 Adobe240 数据集上的比较

由于 Adobe240 数据集没有读出视频时间,因此遵循该数据集的两倍时间超分辨率设置与其他方法进行比较。

定量结果如表 2 所列,在合成数据集 Adobe240 上的实验证明,本文模型能够在当前最新的模型中保持优势地位,同时实现了在低参数量和高运行效率上的突破。FFiT 优于当前的大部分技术,并在参数量和运行效率上优于现有方法。此外,在输入相同的情况下,FFiT 在 5.58 秒内实现了 60 次推理并保持了良好的性能。图 7 给出了本文在 Adobe240 数据集上训练的独立的 FFiT 模型预测的结果。

表 2 在 Adobe240 合成数据集与 RBI 真实数据集上与最新技术的比较

Table 2 Comparison with state-of-the-arts on synthetic dataset Adobe240 and real word dataset RBI

Method	input	Adobe240		RBI		Runtime		Params	FLOPs
		PSNR $\uparrow$	SSIM $\uparrow$	PSNR $\uparrow$	SSIM $\uparrow$	1 $\times$ /s	60 $\times$ /s		
EDVR+XVFI ^[16]	256 $\times$ 256	33.19	0.934	28.17	0.847	0.294	17.64	2.92×10^7	2298×10^9
Jin et al. ^[3]	256 $\times$ 256	32.47	0.924	27.73	0.853	0.250	15.00	1.08×10^7	814×10^9
RPF ₄ ^[4]	256 $\times$ 256	33.32	0.935	28.55	0.872	0.746	44.76	1.14×10^7	1501×10^9
DeMFI ^[5]	256 $\times$ 256	34.34	0.945	29.03	0.895	0.513	30.78	7.41×10^6	1073×10^9
BiT ^[1]	256 $\times$ 256	34.34	0.948	29.9	0.900	0.203	5.76	1.13×10^7	317×10^9
FFiT	256 $\times$ 256	33.87	0.938	29.6	0.897	0.180	5.58	3.584×10^6	169×10^9



图 7 在 Adobe240 数据集上的定性对比

Fig. 7 Qualitative comparison on Adobe240 dataset

4.3.2 在 RBI 数据集上的比较

在 RBI 数据集上,因为在真实的情况中视频存在曝光时间,所以本文在中间的模糊图像上进行比较。本文模型的去除真实模糊的能力优于绝大部分对比模型,并且和最优模型之间的差距微小,同时又在参数量轻量化和运行效率上远优于最新模型。图 8 给出了本文在 RBI 数据集上训练的独立的 FFiT 模型预测的结果。

本文对合成数据集 Adobe240 和真实数据集 RBI 进行了权重交叉评估。首先,展示了在 Adobe240 和 RBI 上训练的独立 FFiT 模型预测的 RBI 样本的结果,在 Adobe240 上训练的模型预测结果中观察到了严重的伪影;相反,在实验中对合成数据进行测试,结果表明在 RBI 上训练的模型不会引入伪影,证明模型在真实模糊中仍然具有很高的鲁棒性。



图 8 在 RBI 数据集上的定性对比

Fig. 8 Qualitative comparison on RBI dataset

4.4 消融研究

本文在 BRI 数据集上进行了消融研究,旨在验证 FFiT

中各模块对模型去模糊性能和模型复杂度的影响程度。使用了残差窗口注意力网络作为骨干网络,输入图像裁剪尺寸为

256×256。

在整个消融实验中采用控制变量法，分别单独将 HQFT, DYRM, MRTB 用于优化 FasterViT，从而验证 FFiT 中各模块对模型去模糊性能和模型复杂度的影响程度。实验结果如表 3 所列。针对不同策略对模型性能的影响，本文进行了如图 9 所示的可视化对比。

表 3 不同 MRTB 参数配置下模型在 RBI 数据集上的表现
Table 3 Performance of model on RBI dataset with different MRTB parameters

	S=2			S=3	
	r=1	r=2	r=3	r=1	r=2
PSNR ↑	29.13	29.32	29.47	29.6	29.67
SSIM ↑	0.889	0.891	0.894	0.897	0.898
Params	2.871×10 ⁶	3.538×10 ⁶	4.205×10 ⁶	3.584×10⁶	4.24×10 ⁶
FLOPs	133.92×10 ⁹	167.26×10 ⁹	200.59×10 ⁹	169.03×10⁹	202.37×10 ⁹

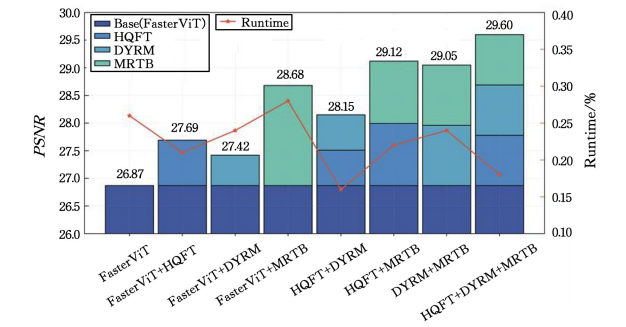


图 9 各个策略对模型的影响
Fig. 9 Impact of each strategy on model

4.4.1 MRTB 分析

表 4 列出了基于 FasterViT 与基于 FasterViT+MRTB 两种模型在去模糊任务上的性能对比。实验数据显示，加入 MRTB 后可同时提升模型的参数表现与运行效率。具体而言，PSNR 提升了 1.81dB, SSIM 提升了 3.5 个百分点，表明本文提出的多尺度残差注意力模块不仅有助于增强模型对运动模糊的处理能力，还能在相同或更少的计算量下带来更稳定的特征表达。

表 4 消融实验结果
Table 4 Ablation study results

基础模型	方法			PSNR	SSIM	Runtime/s
	HQFT	DYRM	MRTB			
FasterViT				26.87	0.842	0.26
	✓			27.69	0.861	0.21
		✓		27.42	0.853	0.24
			✓	28.68	0.877	0.28
	✓	✓		28.15	0.872	0.16
	✓		✓	29.12	0.892	0.22
		✓	✓	29.05	0.887	0.24
	✓	✓	✓	29.60	0.897	0.18

MRTB 选择不同尺度级别 S 和缩放比例 r 时的模型参数及性能对比如表 3 所列。实验结果证明，尺度级别对模型的性能影响更大，同时对参数量的影响较小。相较而言，缩放比例虽然能够增强通道维度的表达能力并实现非线性的映射，但会导致模型的运算量大幅增加。

如图 10 所示，滑动窗口注意力在图像重建的过程中比分层注意力更适合捕捉局部特征与全局依赖之间的相互作用，

而通道重构模块的引入则进一步增强了通道间的信息交流并恢复了在注意力计算过程中被稀释的信息。尽管该改进增加了少量的参数量，但相对于提升的去模糊性能而言，其代价可以忽略不计。这也再次验证了在图像去模糊任务中增强局部注意力与通道特征互通的重要性。

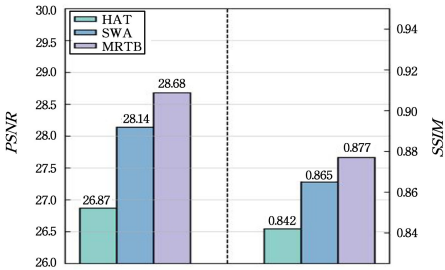


图 10 HAT,SWA,MRTB 在 RBI 数据集上的性能比较
Fig. 10 Performance comparison of HAT,SWA,MRTB on RBI dataset

4.4.2 HQFT 分析

基于 FasterViT 的模型性能与基于 FasterViT+HQFT 的模型性能的比较结果如表 4 所列。可以看到，将 HQFT 模块嵌入网络后，模型的参数指标与运行效率同时提升，在去模糊指标上，PSNR 提升 0.82 dB, SSIM 提升 2 个百分点。这主要归功于 HQFT 所采用的线性操作大幅降低了特征增广的计算难度，同时还能在下采样得到的特征图中保留或凸显更多冗余信息。

与原有的卷积运算与下采样模块相比，HQFT 的线性操作不仅复杂度更低，而且生成的冗余特征图在星型操作的帮助下能够更全面地传递其中的潜在信息。冗余特征的存在使得模糊区域在特征空间中更加突出，进而通过后续星型操作充分利用隐式非线性信息，增强模型对于模糊区域的恢复能力。

4.4.3 DYRM 分析

基于 FasterViT 的模型性能与基于 FasterViT+DYRM 的去模糊与插帧任务上的性能差异如表 4 所列。由表可知，加入 DYRM 后，模型的参数指标与运行效率得以强化，PSNR 提高 0.55 dB, SSIM 提高 1.1 个百分点。与传统的上采样方法相比，DYRM 更注重动态更新点采样集，从而实现了在不增加卷积计算的前提下提高上采样灵活度与精度。

DYRM 通过计算坐标偏移量 offset 的方式，令网络在插值过程中自适应地匹配不同像素位置的运动程度与模糊程度，从而以更低的参数量获取更高的细节还原度。此动态上采样方法为去模糊重建提供了额外的可调节自由度，既能保证全局结构的连贯性，又能对局部区域的细节进行有效修复。

结束语 本文针对运动图像去模糊，改进了 CNN 和 Transformer 的混合框架，提出了新颖且高效的 FFiT 模型，可在任意时间点上对运动图像进行模糊插帧，在小参数量和低计算复杂度上达到了先进水平。该方法通过多尺度残差网络与滑动窗口注意力机制的结合，显著提升了模型的处理精度与鲁棒性，在参数量极为精简的前提下，依旧能实现领先的去模糊与插帧效果。此外，HQFT 模块将 CNN 与 Transformer 的特征表征方式进行更紧密的衔接，使深层的非线性

隐式特征能够无损且高效地在网络中传播,并借助轻量化的线性操作增强了模型的泛化能力。最后,DYRM 采用动态更新采样点的方法,在真实复杂模糊场景中依然保持了出色的稳定性。

当前模型主要依靠离散时间点的监督信号,在运动速度极快的场景下可能仍不足以完全捕捉所有时序细节。随着新特征传递模块研究的深入,CNN 与 Transformer 架构仍蕴含更多潜能,有望在更深层连接或更多维度的融合中实现更优的性能与更低的参数量,并实现在移动/嵌入式设备上的应用。未来将挖掘 CNN+Transformer 框架在运动模糊插帧场景下的潜力,对极端运动情况进行更全面的研究;探索隐式特征间的关联性,精细挖掘非线性隐形特征在提升极端模糊恢复能力中的作用,为复杂运动场景的视频去模糊与插帧提供更为强大的技术支持。

参 考 文 献

- [1] ZHONG Z, CAO M, JI X, et al. Blur interpolation transformer for real-world motion from blur[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023;5713-5723.
- [2] KIM T, LEE J, WANG L, et al. Event-guided deblurring of unknown exposure time videos[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022; 519-538.
- [3] JIN M, HU Z, FAVARO P. Learning to extract flawless slow motion from blurry videos[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019; 8112-8121.
- [4] SHEN W, BAO W, ZHAI G, et al. Video frame interpolation and enhancement via pyramid recurrent framework[J]. IEEE Transactions on Image Processing, 2020, 30: 277-292.
- [5] OH J, KIM M. Demfi: deep joint deblurring and multi-frame interpolation with flow-guided attentive correlation and recursive boosting[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022; 198-215.
- [6] ROZUMNYI D, OSWALD M R, FERRARI V, et al. Motion-from-blur: 3d shape and motion estimation of motion-blurred objects in videos[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022; 15990-15999.
- [7] GUPTA A, AICH A, ROY-CHOWDHURY A K. Alanet: Adaptive latent attention network for joint video deblurring and interpolation[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020; 256-264.
- [8] 杨雪松,何亮田.基于即插即用的盲二值图像去模糊算法[J].重庆工商大学学报(自然科学版), 2025, 42(4): 36-43.
- [9] SHANG W, REN D, YANG Y, et al. Joint Video Multi-Frame Interpolation and Deblurring under Unknown Exposure Time[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023; 13935-13944.
- [10] CHO S J, JIS W, HONG J P, et al. Rethinking coarse-to-fine approach in single image deblurring[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021; 4641-4650.
- [11] PARK D, KANG D U, KIM J, et al. Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training[C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2020; 327-343.
- [12] REN D, ZHANG K, WANG Q, et al. Neural blind deconvolution using deep priors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020; 3341-3350.
- [13] ZHONG Z, GAO Y, ZHENG Y, et al. Efficient spatio-temporal recurrent neural network for video deblurring[C]//Computer Vision-ECCV 2020; 16th European Conference. Springer International Publishing, 2020; 191-207.
- [14] PAN J, BAI H, TANG J. Cascaded deep video deblurring using temporal sharpness prior[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020; 3043-3051.
- [15] SHANG W, REN D, ZOU D, et al. Bringing events into video deblurring with non-consecutively blurry frames[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021; 4531-4540.
- [16] WANG X, CHAN K C K, YU K, et al. Edvr: Video restoration with enhanced deformable convolutional networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019; 1954-1963.
- [17] ZHAO Q, ZHOU D M, YANG H, et al. Image deblurring based on residual attention and multi-feature fusion[J]. Computer Science, 2023, 50(1): 147-155.
- [18] ZAMIR S W, ARORA A, KHAN S, et al. Multi-stage progressive image restoration[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021; 14821-14831.
- [19] ZAMIR S W, ARORA A, KHAN S, et al. Restormer: Efficient transformer for high-resolution image restoration[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022; 5728-5739.
- [20] LIANG J, FAN Y, XIANG X, et al. Recurrent video restoration transformer with guided deformable attention[J]. Advances in Neural Information Processing Systems, 2022, 35: 378-393.
- [21] CAO M, FAN Y, ZHANG Y, et al. VDTR: Video deblurring with transformer[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 33(1): 160-171.
- [22] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention[C]//International Conference on Machine Learning. PMLR, 2021; 10347-10357.
- [23] GRAHAM B, EL-NOUBY A, TOUVRON H, et al. Levit: a vision transformer in convnet's clothing for faster inference[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021; 12259-12269.
- [24] WANG W, XIE E, LI X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions

[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:568-578.

[25] YUAN L, CHEN Y, WANG T, et al. Tokens-to-token vit: Training vision transformers from scratch on imagenet[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:558-567.

[26] HATAMIZADEH A, YIN H, HEINRICH G, et al. Global context vision transformers[C]// International Conference on Machine Learning. PMLR, 2023:12633-12646.

[27] ZHANG P, DAI X, YANG J, et al. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:2998-3008.

[28] YUAN L, HOU Q, JIANG Z, et al. Volo: Vision outlooker for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(5):6575-6586.

[29] LIU Z, HU H, LIN Y, et al. Swin transformer v2: Scaling up capacity and resolution[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 12009-12019.

[30] CHEN Z, XIE L, NIU J, et al. Visformer: The vision-friendly transformer[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:589-598.

[31] CHEN C F R, FAN Q, PANDA R. Crossvit: Cross-attention multi-scale vision transformer for image classification[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:357-366.

[32] LIU X F, JIANG M R, HUANG Y Q, et al. Using image stabilization and VLBE algorithm to extract foreground target from jitter video[J]. Computer Science, 2022, 49(S2):404-411.

[33] BAO W, LAI W S, MA C, et al. Depth-aware video frame interpolation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:3703-3712.

[34] LEE H, KIM T, CHUNG T, et al. Adacof: Adaptive collaboration of flows for video frame interpolation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:5316-5325.

[35] XU X, SIYAO L, SUN W, et al. Quadratic video interpolation [C]// NeurIPS 2019. 2020:1636-1645.

[36] CHI Z, MOHAMMADI NASIRI R, LIU Z, et al. All at once: Temporally adaptive multi-frame interpolation with advanced motion modeling[C]// Computer Vision-ECCV 2020: 16th European Conference. 2020:107-123.

[37] PUROHIT K, SHAH A, RAJAGOPALAN A N. Bringing alive blurred moments[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:6830-6839.

[38] ARGAW D M, KIM J, RAMEAU F, et al. Motion-blurred video interpolation and extrapolation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021:901-910.

[39] HATAMIZADEH A, HEINRICH G, YIN H, et al. Fastervit: Fast vision transformers with hierarchical attention[J]. arXiv: 2306.06189, 2023.

[40] HAN Q, FAN Z, DAI Q, et al. On the connection between local attention and dynamic depth-wise convolution[J]. arXiv: 2106.04263, 2021.

[41] SU S, DELBRACIO M, WANG J, et al. Deep video deblurring for hand-held cameras[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:1279-1288.



CHENG Zhirong, born in 2000, post-graduate. His main research interests include CNN+Transformer and motion video deblur.



XU Yang, born in 1980, associate professor, is a member of CCF (No. D7127M). His main research interests include deep learning and action gesture recognition.

(责任编辑:何杨)