自动化学报
ACTA AUTOMATICA SINICA

基于深度学习的视频插帧研究进展

吴晨阳 张勇 韩树豪 郭春乐 李重仪 程明明

Research Advances on Deep-learning Based Video Frame Interpolation

WU Chen-Yang, ZHANG Yong, HAN Shu-Hao, GUO Chun-Le, LI Chong-Yi, CHENG Ming-Ming

在线阅读 View online: https://doi.org/10.16383/j.aas.c240572

基于深度学习的视频插帧研究进展

吴晨阳 [1] 张 勇 [1, 2] 韩树豪 [1] 郭春乐 [1, 3] 李重仪 [1, 3] 程明明 [1, 3]

摘 要 视频插帧技术是视频处理领域的研究热点问题. 它通过生成中间帧来提高视频的帧率, 从而使视频播放更加流畅, 在老视频修复、电影后期制作和慢动作生成等领域发挥着重要的作用. 随着深度学习技术的迅猛发展, 基于深度学习的视频插帧技术已经成为主流. 本文全面综述现有的基于深度学习的视频插帧工作, 并且深入分析这些方法的优点与不足. 随后, 详细介绍视频插帧领域的常用数据集, 这些数据集为视频插帧相关研究和算法训练提供重要支撑. 最后, 对当前视频插帧研究中仍然存在的挑战进行深入思考, 并且从多个角度展望未来的研究方向, 旨在为该领域后续的发展提供参考.

关键词 视频插帧, 深度神经网络, 卷积神经网络

引用格式 吴晨阳, 张勇, 韩树豪, 郭春乐, 李重仪, 程明明. 基于深度学习的视频插帧研究进展. 自动化学报, 2025, **51**(8): 1−17

DOI 10.16383/j.aas.c240572 **CSTR** 32138.14.j.aas.c240572

Research Advances on Deep-learning Based Video Frame Interpolation

WU Chen-Yang[1] ZHANG Yong[1, 2] HAN Shu-Hao[1] GUO Chun-Le[1, 3]
LI Chong-Yi[1, 3] CHENG Ming-Ming[1, 3]

Abstract Video frame interpolation technology has become a hot research topic in the field of video processing. It improves the frame rate of videos by generating intermediate frames, thereby making video playback smoother. It plays a crucial role in various fields such as old video restoration, film post-production and slow-motion generation. With the rapid development of deep learning technology, video frame interpolation technology based on deep learning has become mainstream. This paper comprehensively reviews the existing deep-learning based video frame interpolation works and deeply analyzes the advantages and disadvantages of these methods. Subsequently, this paper elaborately introduces the commonly used datasets in the field of video frame interpolation. These datasets provide important support for video frame interpolation-related research and algorithm training. Finally, the paper deeply contemplates the challenges existing in current video frame interpolation research and looks ahead to future research directions from multiple perspectives, aiming to provide references for the subsequent development of this field.

Key words Video frame interpolation, deep neural networks, convolutional neural networks

Citation Wu Chen-Yang, Zhang Yong, Han Shu-Hao, Guo Chun-Le, Li Chong-Yi, Cheng Ming-Ming. Research advances on deep-learning based video frame interpolation. *Acta Automatica Sinica*, 2025, **51**(8): 1−17

在视频拍摄或传输中, 由于摄像机性能、网络带宽、视频编码器和存储空间等因素的限制, 往往会导致获取到的视频存在帧率下降的问题. 帧率的下降不仅直接降低视频的视觉质量, 还严重影响用户的观看体验. 例如, 视频的低帧率导致的画面不流畅、观看舒适度降低等. 因此, 如何提高已有视频的帧率, 提升视频质量, 一直以来都是视频处理技术领域中亟须解决的问题.

为显著提升视频质量并解决播放不流畅的问题, 视频插帧 (Video frame interpolation, VFI) 技术被提出. 该技术旨在现有视频序列中生成新的帧, 以提高视频帧率, 进而实现更流畅、更平滑的视频播放效果, 并显著改善视觉体验. 视频插帧的研究包含图像处理、计算机视觉及优化理论等领域的核心问题, 例如图像特征提取、运动估计和最优化算法等. 视频插帧不仅作为这些基础问题的一个实际

收稿日期 2024-08-14 录用日期 2025-04-10
Manuscript received August 14, 2024; accepted April 10, 2025
国家自然科学基金 (62306153, U23B2011, 62176130), 中央高校基本科研业务费 (070-63243143), 天津市自然科学基金 (24JCJQJC00020), 深圳市科技计划 (JCYJ20240813114237048) 资助
Supported by National Natural Science Foundation of China (62306153, U23B2011, 62176130), Fundamental Research Funds for the Central Universities (070-63243143), Natural Science Foundation of Tianjin (24JCJQJC00020) and Shenzhen Science and Technology Program (JCYJ20240813114237048)
本文责任编委 丛杨
Recommended by Associate Editor CONG Yang
1. 南开大学计算机学院 天津 300350 2. 重庆长安望江工业集团有限公司 重庆 404100 3. 南开国际先进研究院 (深圳福田) 深圳 518048
1. College of Computer Science, Nankai University, Tianjin 300350 2. Chongqing Chang′an Wangjiang Industry Co., Ltd., Chongqing 404100 3. Nankai International Advanced Research Institute (SHENZHEN-FUTIAN), Shenzhen 518048

应用场景, 同时也促进相关领域的发展. 因此, 视频插帧技术不仅在学术界引起广泛关注, 更成为计算机视觉等相关领域研究的热门方向, 其研究价值和实用意义不容忽视.

视频插帧算法的研究起源可以追溯到 20 世纪 80 年代末到 90 年代初. 随着计算机视觉和图像处理领域的发展, 人们开始探索如何通过技术手段改进视频质量, 特别是如何在不增加视频原始拍摄成本的情况下提高视频播放的帧率.

最初, 视频插帧技术主要依靠传统的图像处理技术, 如帧间差分、光流估计等方法来合成中间帧. 随着时间的推移, 特别是进入 21 世纪以来, 深度学习技术的发展极大地推动视频插帧技术的发展. 从 2015 年开始, 利用深度神经网络进行视频插帧的研究逐渐增多, 研究者开始尝试使用卷积神经网络、Transformer 和扩散模型 (Diffusion model) 等深度学习模型来自动学习视频帧之间的运动和变化规律, 从而更准确、更自然地生成中间帧.

视频插帧技术的应用范围极为广泛, 涵盖老旧视频的修复、虚拟现实 (Virtual reality, VR)、增强现实 (Augmented reality, AR)、动画及电影制作以及视频慢动作生成等多个领域. 特别值得一提的是, 在生成式人工智能 (Artificial intelligence generated content, AIGC) 迅猛发展的当下, 以 Sora 为代表的视频生成技术对社会生产与生活产生较大的影响. 为制作更加流畅的视频内容, 这类视频生成技术通常只生成一段视频的关键帧, 随后使用视频插帧技术来实现视频流的平滑化. 这不仅展现视频插帧技术在支持 AIGC 视频创作方面的关键作用, 也突显其在当前人工智能技术快速进步背景下的重要应用价值.

在 2015 年之前, 视频插帧算法的研究多集中在如何对光流进行准确的估计. 如图 1 所示, 随着深度学习技术的迅猛发展, 基于深度学习的视频插帧算法分为基于相位的方法、基于核的方法、基于生成的方法和基于光流的方法, 各个方法的提出和发展均推动视频插帧技术的进步. 基于相位的方法不需要显式的运动估计, 通过简单的像素级相位调整能够快速生成中间帧. 基于核的方法通过学习卷积核的知识来捕捉运动信息, 能够有效处理遮挡和动态模糊. 基于生成的方法能够直接预测中间帧, 在复杂运动和光照变化的场景中表现出明显的优势. 基于光流的方法能精准建模视频运动, 成为当前视频插帧领域的研究热点, 也被认为是最具潜力的方向. 根据光流计算方式的不同, 基于光流的方法可分为基于前向翘曲的视频插帧方法和基于后向翘曲的视频插帧方法. 前者能更加显式地建模物体运动, 但可能导致空洞问题; 而后者通过隐式建模避免了空洞问题, 但在某些应用场景中性能会有所下降. 本文将重点介绍基于深度学习的视频插帧技术及其典型方法. 第 1 节详细梳理目前主流的基于深度学习的视频插帧技术; 第 2 节介绍视频插帧中常用的

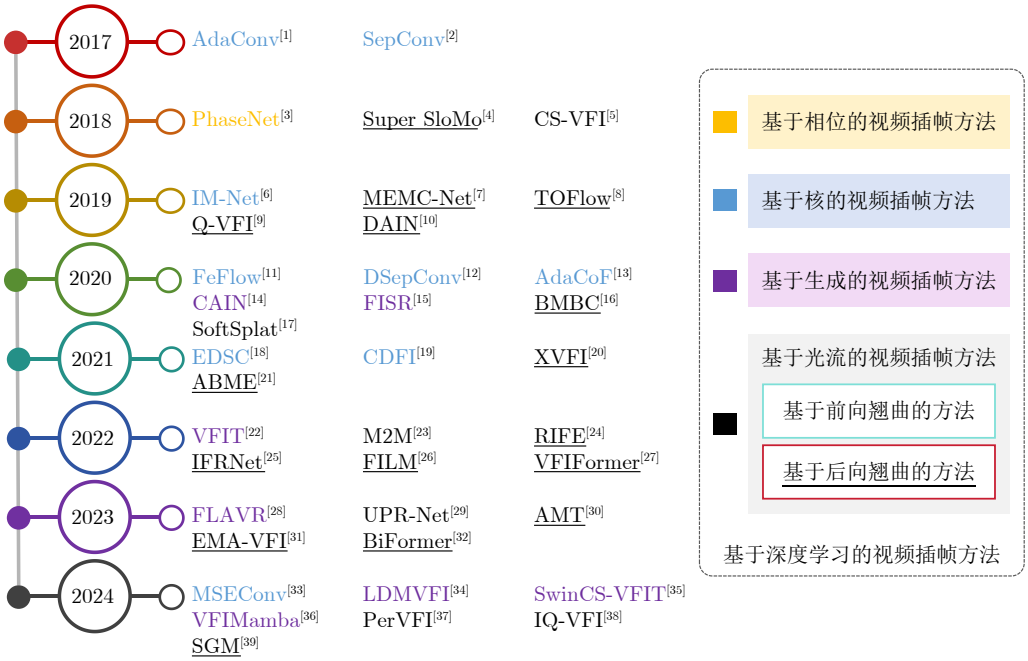


图 1 基于深度学习的视频插帧发展流程图

Fig. 1 The flowchart of deep-learning based video frame interpolation development

数据集; 最后对该方向存在的问题进行思考和展望, 并总结本文的研究内容.

近年来对视频插帧技术的研究热情更为高涨, 近三年 (2022 年至 2024 年) 提出的视频插帧算法数量超过本文统计算法总数的 47%. 与文献 [40−41] 对视频插帧领域进行的综述相比, 本文在统计范围和分析视角方面存在显著区别. 在统计范围方面, 现有综述统计截止于 2021 年, 而本文的统计时间跨度覆盖到 2024 年, 全面纳入近三年来涌现的新算法与研究成果. 这使得本文不仅涵盖前人综述中的内容, 还囊括后续三年间该领域的重要发展, 例如, 采用 Transformer 架构的 VFIT[22]、实时轻量化的插帧网络 RIFE[24] 和使用扩散模型进行帧预测的 LDMVFI[34], 能够更全面地展现视频插帧技术的发展脉络和最新动态. 在分析视角方面, 本文深入剖析在大模型快速发展的背景下视频插帧技术面临的挑战与机遇, 而且针对多种插帧挑战提出更为深刻的思考, 为推动该技术的进一步发展提供新的视角.

1 基于深度学习的视频插帧方法

近年来, 深度学习受到广泛关注, 深度学习的概念源于尝试模拟人脑进行分析和学习的人工神经网络. 通过多层次的数据抽象和特征提取, 深度学习模型可以识别、分类和预测各种复杂模式. 这使得深度学习在图像分类、目标检测和视频分析等多个领域都显示出巨大的潜力. 研究者们借鉴深度学习在图像分类、语义分割和目标检测中的成功经验, 提出基于深度学习的视频插帧方法.

对于视频插帧领域, 基于深度学习的方法通过训练大量的视频数据, 使模型学习临近帧到中间帧之间的映射关系. 深度学习方法不仅学习两帧图像之间显式的运动模式 (如光流), 还能隐式地捕捉临近帧之间的时间动态和光照等模式的变化. 大部分基于深度学习的视频插帧算法采用端到端的策略, 即直接学习输入帧到输出帧的映射, 无需人为设计特征和预定义运动模型. 因为可以使模型自动从数据中学习到最优的特征表示和转换规则, 所以这种端到端的学习方式可以使算法更好地适应各种复杂的视频内容和运动模式.

通常情况下, 按照处理方法和主要侧重点的不同, 本文将基于深度学习的视频插帧方法分为基于相位的视频插帧方法、基于核的视频插帧方法、基于生成的视频插帧方法和基于光流的视频插帧方法. 基于光流的视频插帧方法通常被分为基于后向翘曲 (Backward warping) 的视频插帧方法和基于前向翘曲 (Forward warping) 的视频插帧方法. 除此之外, 由于基于光流的视频插帧方法中光流估计器的设计较广泛地参考光流估计领域的模型设计, 为增加读者对这部分的了解, 本文在介绍基于光流的视频插帧方法前还简单介绍基于深度学习的光流估计方法.

1.1 基于相位的视频插帧方法

相位 (Phase) 在信号处理、物理学和数学等领域是一个重要概念, 它主要描述波形 (或信号) 周期性变化中的某个特定点的位置, 被应用于计算机视觉的多个领域[3, 42−43]. 在视频或图像处理中, 特别是处理基于时间的信号 (如视频帧) 时, 相位信号被用来捕捉和描述物体的运动. 由于物体的运动会导致物体的亮度模式随时间发生改变, 所以这种变化可以通过分析相位的变化来进行量化[3, 42, 44]. 为揭示和放大视频中几乎不可察觉的运动, Wadhwa 等[44] 提出一种基于相位的视频运动处理技术, 通过对复数值可导向金字塔系数的相位变化在时间上进行处理和放大, 该技术能够更进一步放大较小的运动, 同时产生更少的伪影和噪声.

基于 Wadhwa 等[44] 的工作, 研究人员将基于相位的方法引入视频插帧中, 提出基于相位的视频插帧方法. 基于相位的方法不需要任何形式的显式运动估计, 仅通过进行简单的逐像素相位修改即可求得中间帧. 2015 年, Meyer 等[42] 首先提出基于相位的视频插帧算法. 该算法通过在多尺度金字塔的定向层次中传播相位信息, 来校正相位位移. 同时, 为避免大运动产生的伪影, 该算法提出基于相位位移的自适应上限来限制相位变化的幅度, 从而保证运动的平滑过渡. 对比基于光流的方法, Meyer 等[42] 认为基于相位的方法不会产生严重的视觉伪影, 而且对于光照变化的情况, 性能表现更稳定.

2018 年, 受到以神经网络为代表的深度学习技术研究热潮的影响, Meyer 等[3] 提出 PhaseNet. 基于 Meyer 等[42] 在 2015 年提出的方法, PhaseNet 结合基于相位的方法和深度学习方法, 使用神经网络解码器逐层估计中间帧的相位和振幅值, 然后根据这些值来重构中间帧. PhaseNet 独立处理输入图像的通道, 并在通道和金字塔级别之间共享权重, 这不仅使得其在处理大运动、运动模糊或光照变化等场景中有更好的性能, 而且显著降低了网络的参数量. 值得注意的是, 为提高训练效率和稳定性, PhaseNet 从最粗粒度的尺度分层训练, 并逐步进入下一个更细的尺度.

综上所述, 基于相位的 VFI 方法通过精巧地利用相位信息, 结合传统图像处理技术和先进的深度

学习框架, 为视频插帧任务提供一种高效、稳定且视觉效果良好的解决方案.

1.2 基于核的视频插帧方法

基于核的视频插帧方法的核心思想是利用一组学习得到的核 (Kernel) 对视频帧进行处理, 来估计中间帧. 这种方法紧密依赖深度学习模型, 通过训练学习得到的卷积核来捕捉和模拟视频序列中的运动模式. 更具体的说, 基于核的方法首先分析视频序列中连续帧之间的内容和运动信息, 然后通过应用一组卷积核来预测中间帧的像素值. 这些卷积核能够捕捉到物体的运动轨迹和速度, 以及场景变换的细节, 从而生成自然流畅的中间帧.

2017 年, Niklaus 等[1] 首次提出基于核的视频插帧方法 (也称为 AdaConv), 该方法使用自适应卷积将运动估计和帧合成结合到一个卷积步骤中. 将该方法表达为:

$$out_{x,\,y} = AdaConv(K_{x,\,y},\ [I_0^{x,\,y,\,(p)},\ I_1^{x,\,y,\,(p)}]) \quad (1)$$

其中, $out_{x,\,y}$ 表示位置为$(x,\,y)$ 的最终预测插值结果; $AdaConv(K,\,In)$ 表示使用自适应卷积核 K 对输入 In 的自适应卷积操作; $I_0^{x,\,y,\,(p)}$ 表示第 0 帧的以 $(x,\,y)$ 为中心、p 为半径的区域; $[I_0^{x,\,y,\,(p)},\ I_1^{x,\,y,\,(p)}]$ 作为输入 In 使用卷积核 $K_{x,\,y}$ 进行卷积. 自适应卷积核 $K_{x,\,y}$ 的计算方式可以表示为:

$$K_{x,\,y} = Convs([I_0^{x,\,y,\,(r)},\ I_1^{x,\,y,\,(r)}]),\ r > p \quad (2)$$

其中, $Convs(\cdot)$ 表示可学习的卷积神经网络, 获得自适应空间卷积核 $K_{x,\,y}$ 的方法是对 $[I_0^{x,\,y,\,(r)},\ I_1^{x,\,y,\,(r)}]$ 信息进行卷积. 为保证得到的自适应空间卷积核 $K_{x,\,y}$ 能够捕捉视频序列中的运动模式, 所以要求 $r > p$.

总的来说, 该方法通过利用神经网络的强大学习能力, 将图像的局部运动信息拟合为一个自适应空间卷积核, 然后使用该自适应空间卷积核指导中间帧的合成. 该方法的优势在于处理遮挡、模糊和亮度变化等复杂场景时能够得到较为清晰的结果, 但是这种卷积方式需要估计大量的核来处理大尺度的运动, 不能同时合成高分辨率的视频.

为解决高分辨率视频插帧需求场景带来的巨大内存和算力消耗问题, 基于文献 [1], Niklaus 等[2] 提出使用自适应可分离式卷积的方法 (也称为 SepConv) 来减少内存和算力依赖. 由于使用方法的局限性和模型设计, SepConv 和 AdaConv 都不能处理物体较大尺度的运动. 针对高分辨率 (大尺度运动) 视频插帧, Peleg 等[6] 提出 IM-Net, 其核心贡献在于给出一种结构化的网络架构和端到端训练策略, 值得注意的是, IM-Net 将运动估计定义为分类任务而非回归问题.

针对复杂运动场景的插帧问题, 如遮挡、模糊和亮度突变等挑战, Gui 等[11] 提出 FeatureFlow (也称为 FeFlow), 通过利用深度结构感知特征来预测特征流, 进而生成中间帧的结构图像, 最后进行纹理细化获得中间帧结构. Cheng 等[12] 提出基于可变形可分离式卷积的 DSepConv, 不仅学习空间自适应可分离式卷积核, 还学习可变形的偏移量 (Offset) 和掩膜 (Mask). 特别的是, DSepConv 应用小卷积核来处理大位移运动场景, 在一定程度上获得较好的处理结果. 基于光流和可变形卷积的方法, Lee 等[13] 提出一种新的翘曲 (Warping) 方法 AdaCoF. 该方法强调基于核的视频插帧方法自由度都是受限的, 为实现算法的灵活性, 通过自适应地调整自适应卷积核的形状和大小, 更精准地捕获物体的运动和形变信息, 有利于处理复杂运动和遮挡场景下的视频插帧任务.

在 DSepConv[12] 基础之上, Cheng 等[18] 提出一种增强的可变形可分离式卷积 (Enhanced deformable separable convolution, EDSC) 的视频插帧新方法, 通过将时间步骤作为网络的一个控制变量, 直接生成视频序列中任意时间点的帧, 而不需要采用多步骤或递归的方法, 提高了视频插帧的灵活性和效率. 为降低视频插帧模型的参数量和计算量, Ding 等[19] 大幅度压缩 AdaCoF[13] 模型, 并引入多分辨率的翘曲模块提出 CDFI 模型[19]. CDFI 模型仅使用 AdaCoF 四分之一的参数量和计算量, 其性能超过了 AdaCoF 的性能. Ding 等[33] 针对视频插帧任务中复杂运动建模的问题, 提出一种统一的变形卷积框架——多尺度可扩展变形卷积 (Multi-scale expandable deformable convolution, MSEConv), 以同时实现复杂运动建模与帧插值. 该方法通过具有全局注意力的深度全卷积神经网络, 估计具有不同扩展程度的小尺度运动核权重, 并为每个像素的合成自适应地分配权重.

总的来说, 基于核的视频插帧方法在处理复杂运动场景时有着高度的灵活性和自适应能力, 同时利用基于深度学习的端到端训练方式, 简化了插帧过程, 无需复杂的预处理和后处理步骤. 但是基于核的方法在处理高分辨率视频时对内存和算力的需求比较高. 同时, 基于核的方法仅能够处理一定范围内的运动, 在处理超过特定像素距离的运动时, 性能可能会下降.

1.3 基于生成的视频插帧方法

在一些研究中, 将该类方法[25, 30] 写为“Hallucination-based” (直译为: 基于幻想的方法), 考虑到这

类方法的核心实现机制, 将其称之为“基于生成的视频插帧方法”. 这种方法利用深度学习技术来直接产生中间帧, 而非依赖于传统的运动估计手段 (例如光流) 进行间接计算. 通过对大量视频数据的训练, 这类方法能够深入学习视频帧之间的内在联系和规律, 使得模型能够基于两帧输入直接预测并生成一个或多个中间帧, 从而实现平滑的帧间过渡.

2020 年, Choi 等[14] 提出一个神经网络 CAIN, 不需要额外的运动估计模块, 可以进行端到端的训练. CAIN 通过一个特殊的特征重塑操作 (即像素洗牌, PixelShuffle), 结合通道注意力 (Channel attention) 替代传统的光流计算模块. CAIN 方法的主要思想是将特征图中的信息分布到多个通道中, 并通过关注通道来提取像素级帧合成的运动信息. CAIN 可以较好地处理遮挡等复杂场景, 但是其不能显式地捕获输入帧之间的复杂时空依赖关系. 3D 时空卷积在捕获复杂的空间和时间依赖关系方面已经被证明是有效的[45], Kalluri 等[28] 提出 FLAVR, 通过引入 3D 卷积来增强模型对运动的建模能力. 同时, 为方便模型的部署, FLAVR 在输出质量和推理时间上做平衡, 较大程度上降低了计算开销. 针对运动模糊问题, 林传健等[46] 提出一种针对运动模糊问题的新型视频插帧方法, 通过多任务融合卷积神经网络的去模糊和插帧模块, 显著提升了模糊视频的帧率转换效果, 通过端到端计算, 该方法能将两个模糊的视频帧转换为清晰且连续的视频帧. 针对高分辨率视频流, Kim 等[15] 面向视频插帧和超分领域提出 FISR 框架, 该框架采用一个创新的训练策略, 能够处理多个连续的视频帧样本, 并引入新的时间损失来在这些连续样本上施加时间正则化, 以实现更稳定的中间帧生成. 值得注意的是, FISR 还是一个 VFI-SR 的联合框架, 能够同时提高视频的分辨率和帧率 (例如, 将一个 2 K-30 FPS 的视频转换为 4 K-60 FPS 的视频).

随着 Transformer 在计算机视觉的多个领域达到前所未有的优异性能, Shi 等[22] 创新性地将 Transformer 架构应用于视频插帧任务中, 提出视频插帧 Transformer VFIT. 其设计目标是解决深度卷积神经网络在视频插帧中遇到的诸多挑战, 特别是有限的感受野等问题. 为理解和预测复杂视频序列中的动态变化, VFIT 通过自注意力机制精细处理提取的多尺度深层特征, 这可以使模型更容易捕获长距离的内容依赖关系. 针对现实场景中广泛存在的大运动和复杂运动问题, 石昌通等[35] 提出 SwinCS-VFIT 方法, 这是一种利用改进 ViT 以提升视频插帧模型性能的方法. SwinCS-VFIT 通过融合跨尺度窗口注意力和可分离式时空注意力, 增大了模型的感受野. SwinCS-VFIT 还对时空依赖以及远程像素依赖关系进行建模, 以增强模型对较大运动场景的处理能力. 基于 Mamba 在各个任务上的优秀表现, Zhang 等[36] 提出 VFIMamba, 一种基于选择性状态空间模型 (Selective state space models, S6) 的新型视频插帧方法, 通过高效动态的帧间建模显著提升了插帧效果. 该方法设计混合状态空间模块 (Mixed-SSM block, MSB), 通过交错帧令牌重组和多方向 S6 建模实现了高效的信息传递, 并保持线性复杂度. 基于事件相机拍摄的数据, Cho 等[47] 针对利用事件相机实现视频帧插值的问题, 提出一种测试时自适应方法 TTA-EVF, 旨在解决源域与目标域之间的分布差异. TTA-EVF 在仅提供低帧率视频的目标域上, 通过在线序列学习, 利用可信像素作为伪真值, 实现了稳定且准确的学习.

值得注意的是, 扩散模型在图像生成领域展现出较高的性能, 受此启发, Danier 等[34] 将其应用于视频插帧任务, 并创新性地提出一种基于潜在扩散模型 (Latent diffusion model, LDM) 的方法, 称之为 LDMVFI. 值得注意的是, LDMVFI 指出传统的 VFI 方法大多依赖于深度神经网络, 通过最小化 L_1 损失、L_2 损失或 VGG 损失等常见的优化目标进行训练. 然而, 这些优化指标在感知质量上的表现较差, 尤其在面对复杂运动或动态纹理的场景时, 虽然能够取得较高的峰值信噪比 (Peak signal to noise ratio, PSNR) 等客观指标, 但感知效果往往不理想. 为解决这一问题, LDMVFI 使用潜在扩散模型作为生成框架, 将 VFI 问题形式化为条件生成问题, 这是首次将其应用于视频插帧任务. 具体而言, LDMVFI 结合自编码器和反向扩散过程, 其中自编码器将视频帧映射到隐空间, 反向扩散过程则在隐空间中逐步去噪生成中间帧. LDMVFI 是首次尝试利用 LDM 来解决视频插帧问题的方法, 通过在一个紧凑的潜在空间 (Latent space) 内执行前向和逆向扩散过程来产生高质量的中间帧, LDMVFI 展现出生成高感知质量视频内容的能力.

总的来说, 基于生成的视频插帧方法对于复杂场景有着较强的适应性, 能够较好地处理复杂运动、遮挡以及光照变化等场景. 然而, 在处理那些模型未能充分掌握的复杂动态或纹理模式时, 可能会出现不符合真实情况的细节或伪影问题. 扩散模型在视频插帧任务中展现出广阔的发展前景. 随着生成模型的持续进步, 扩散模型已经在图像生成、视频生成等多个领域展现出卓越的能力, 尤其在提升感知质量方面相较于传统方法具有显著优势.

1.4 基于光流的视频插帧方法

得益于光流的鲁棒性和显式建模视频中运动的能力[48−50], 基于光流的技术已成为基于深度学习的视频插帧技术研究的核心. 这类方法通常分为两个主要阶段: 光流估计和帧合成. 在光流估计阶段, 模型通过分析输入图像或其特征 (即图像中各像素点随时间的运动) 来预测光流. 紧接着, 在帧合成阶段, 利用预测的光流通过翘曲将两个输入帧转换至中间帧的位置, 从而生成所需的中间帧. 这一过程不仅准确地捕捉视频内容的运动, 而且有效地实现流畅的插帧效果.

光流可以显式建模物体的运动, 对于一个物体 A 从 0 时刻 (第 0 帧) 到 1 时刻 (第 1 帧) 的位移矢量可以表示为: $f_{0\to1}$. 一般情况下, 如果要获得中间时刻 (0.5 时刻) 物体 A 的位置, 则可以通过如下的方式计算:

$$P_A^{0.5} = P_A^0 + 0.5 \times f_{0\to1} \tag{3}$$

其中, P_A^0 表示物体 A 原始 (0 时刻) 的位置; $P_A^{0.5}$ 表示中间时刻 (0.5 时刻) 物体 A 的位置, 这个过程也被称作为翘曲.

然而, 这种方法未能充分考虑实际光流与特定任务目标之间存在的偏差, 导致估计出的光流对于某些特定任务而言可能并非最佳选择. 特别是在匀加速、变加速或曲线运动等复杂运动场景中, 仅通过计算中间位置的方法往往难以达到理想的效果, 因为在 0.5 时刻, 物体并不一定恰好位于位移的中间点. TOFlow[8] 首先提出面向任务的光流的概念, 通过学习的方法让模型预测出“虚假”的光流 $f'_{0\to1}$, 使模型能够准确地通过式 (3) 计算物体 A 的新位置, 这种思想为视频插帧技术的发展奠定了基础. 由于光流 $f'_{0\to1}$ 不能代表物体真实的运动位移 $f_{0\to1}$, 只能让模型的推理结果更加逼近真实的中间帧结果, 所以称 $f'_{0\to1}$ 为面向任务的光流.

根据得到最后的中间帧所使用的翘曲方式的不同, 本文将基于光流的视频插帧方法分为基于前向翘曲的视频插帧方法和基于后向翘曲的视频插帧方法, 见文献 [17] 的图 2. 前向翘曲是一种从源图像空间到目标图像空间的映射方法, 由于在计算上常用“抛雪球”算法实现, 所以又称其为喷溅 (Splatting). 前向翘曲在概念上比较直观, 因为并非所有目标像素位置都能从源帧中直接获得对应的像素值, 所以它可能导致目标图像出现空洞的问题. 后向翘曲则是从目标图像空间到源图像空间的映射方法, 由于在算法角度通常对原图像进行采样来实现, 所以又称为采样 (Sampling). 这种方法的一个关键优势是可以直接在输出帧的像素位置进行计算, 避免了空洞 (无像素覆盖的区域) 和重叠 (多个像素映射到同一位置) 的问题.

特别地, 视频插帧模型中的光流估计模块设计通常借鉴基于深度学习的光流估计方法, 以提升光流的精度和效果. 为帮助读者更好地理解视频插帧算法中光流估计模块的作用, 本节首先简要介绍基于深度学习的光流估计方法, 随后详细探讨基于后向翘曲和前向翘曲的两种视频插帧方法.

1.4.1 基于深度学习的光流估计方法

基于深度学习的光流估计方法在视频处理领域中的应用极为广泛, 涵盖视频插帧、运动估计、视频超分等多个重要方向. 其光流估计网络的设计思路不仅为视频插帧领域的光流估计模块设计提供宝贵参考, 而且对视频插帧技术的发展产生深远影响.

2015 年, Dosovitskiy 等[48] 提出的 FlowNet 是第一个使用深度学习技术估计光流的神经网络, FlowNet 包含两个模型 FlowNetS 和 FlowNetCorr. 在 FlowNetCorr 网络中, 作者首次提出计算相关性的概念, 相关性的计算方式被定义为:

$$c(x_1,\ x_2) = \sum_o \langle f_1(x_1+o),\ f_2(x_2+o)\rangle \tag{4}$$

其中, o 满足 $o \in [-k,\ k] \times [-k,\ k]$, k 表示相关性计算的范围; f_1 和 f_2 表示两个相邻帧的特征图; $f_1(x_1+o)$ 表示特征图 f_1 在 (x_1+o) 位置的特征向量. 由于将位置 x_1 和 x_2 全部遍历得出 $c(x_1,\ x_2)$ 会导致计算复杂度的提高, 所以作者限制 x_1 和 x_2 之间的距离不超过某个距离 d 来降低计算复杂度. 由于计算相关性可以帮助模型更好地匹配两帧的目标, 从而更好地解析物体运动, 所以后面的插帧网络较为广泛地采用这种思想来进行设计.

2017 年, Ilg 等[49] 进一步提出 FlowNet 2.0. 通过串联并堆叠多个 FlowNetS 和 FlowNetCorr 模块, 实现对光流的迭代细化处理. 接下来的一系列创新工作, 包括 SpyNet[50]、PWC-Net[51]、LiteFlowNet[52] 以及 VCN[53], 均采用从粗糙到精细的金字塔策略对光流进行迭代优化. 这种逐步细化光流估计的思想已经被公认为光流估计领域的最佳选择, 对视频插帧网络中光流估计模块的设计产生持久而深远的影响. 不同于式 (4) 中 FlowNet 计算相关性的方式, PWC-Net[51] 提出另外一种计算相关性的方式 (即, Cost volume), 其计算方式被定义为:

$$cv(x_1,\ x_2) = \frac{1}{N}({f_1}^{\mathrm{T}}(x_1) f_2(x_2)) \tag{5}$$

其中, $cv(x_1,\ x_2)$ 表示特征向量 $f_1(x_1)$ 和 $f_2(x_2)$ 的 Cost volume 相关性, 与 FlowNetCorr 的思想类似,

Cost volume 方法也限制 $|x_1 - x_2|_\infty \leq d$. 与 FlowNetCorr 方式不同的是, Cost volume 方式是将 $f_1(x_1)$ 的单个特征向量与 f_2 特征图中对应位置邻域的 d^2 个特征向量求相关性.

2020 年, Teed 和 Deng 提出另外一种计算相关性的方法 RAFT[54], 与式 (4) 和 (5) 中计算相关性方法不同的是, 式 (4) 和 (5) 产生的相关性矩阵的大小是 $W \times H \times d^2$ (其中, W 和 H 分别表示特征图的宽和高), 而 RAFT 计算的相关性矩阵大小是 $W \times H \times W \times H$, 计算方式是:

$$C_{ijkl} = {f_1}^{\mathrm{T}}(i,\ j) f_2(k,\ l) \tag{6}$$

其中, $f_1(i,\ j)$ 表示特征图 f_1 在 $(i,\ j)$ 位置的特征向量. 根据式 (6) 不难看出, RAFT 计算相关性的方式是对特征图 (大小为原图像的 1/8) 进行全局相似性计算. 与 SpyNet[50]、PWC-Net[51]、LiteFlowNet[52] 以及 VCN[53] 等网络设计不同的是, RAFT 使用修改后的 GRU (Gate recurrent unit) 模块来细化光流, 这种设计可以使 RAFT 在每一层细化光流时采用同一组参数, 从而降低网络的参数量. 值得注意的是, 受 RAFT 启发, Zhao 等[55] 提出 GMFlowNet, 使用全局匹配 (Global matching) 的方法获得初始化光流, 提高了光流估计的准确性和鲁棒性.

基于深度学习的光流估计算法的思想持续影响基于光流的视频插帧算法的研究. 其中, 逐级细化光流的策略已成为设计基于光流的视频插帧模型的核心要素. 此外, 各类相关性求解方法也被广泛应用于视频插帧网络的设计之中, 例如 AMT[30]、UPR-Net[29] 等 SOTA 模型就采纳此前讨论的相关性算法.

1.4.2　基于后向翘曲的视频插帧方法

由于后向翘曲使用采样法实现, 所以后向翘曲对于隐式运动建模比前向翘曲更具有优势. 而且, 在反向传播算法计算后向传播的梯度时, 后向翘曲更容易实现. 在深度学习越来越受欢迎的今天, PyTorch 成为最受欢迎的深度学习研究工具, 由于 PyTorch 库中 grid_sample(·) 函数的实现, 后向翘曲的并行实现起来更加简单, 而前向翘曲常要借助 Cupy 库来编写 CUDA 内核代码来实现并行.

2018 年, Jiang 等[4] 首次提出使用深度学习来实现视频插帧的网络 Super SloMo, 首先使用 UNet 来计算输入图像之间的双向光流, 其次对其细化, 得到双边光流 $F_{t\to 0}$、$F_{t\to 1}$, 最后根据式 (7) 得到最终帧 I_t.

$$I_t = \frac{(1-t)V_{t\leftarrow 0} \odot g(I_0,\ F_{t\to 0}) + tV_{t\leftarrow 1} \odot g(I_1,\ F_{t\to 1})}{(1-t)V_{t\leftarrow 0} + tV_{t\leftarrow 1}} \tag{7}$$

其中, $g(\cdot,\cdot)$ 表示后向翘曲函数; $V_{t\leftarrow 0}$ 目的是处理遮挡问题, 表示某一区域的像素是否可见; I_0 表示插帧的第 0 帧; I_1 表示插帧的第 1 帧; $\odot$ 表示点乘操作. 使用遮挡编码 V 是 Super SloMo 方法的亮点, 对后续插帧的网络设计产生深远的影响. 结合运动估计和运动补偿方法的优点, Bao 等[7] 提出 MEMC-Net, 对参考帧同时预测光流和补偿核, 最后通过设计的“自适应翘曲”算法来实现中间帧的合成. 由于物体实际运动的复杂性, Xu 等[9] 使用二次曲线来模拟更加复杂的运动, 称其方法为 Q-VFI. 为提升光流估计的质量和有效改善遮挡场景下的插帧效果, 张倩等[56] 提出一种结合光流估计与深度学习的视频插帧方法, 通过端到端卷积神经网络联合建模运动估计和遮挡处理, 利用改进的双向光流和多种加权损失函数生成逼真且高质量的中间帧, 有效提升了插帧效果. 借鉴 PWC-Net 中计算特征相似度来细化光流的思想, Park 等[16] 提出 BMBC, 使用计算双边相似度的方法来获得更准确的双边运动估计, 以实现更加准确的插帧. 在保证插帧模型性能的基础上, RIFE[24] 引入面向任务的蒸馏损失, 有效降低了插帧模型的计算量和参数量, 实现对数据的实时处理. 针对高分辨率视频, Sim 等[20] 提出 XVFI, 其基于递归多尺度共享结构, 由两个级联模块构成, 用于输入两帧之间的双向光流预测. 对于高分辨率数据 (如 4 K 视频等), XVFI 可以较好地捕获非常大的运动和复杂纹理对象的基本信息. 对于如何估计双边光流, Park 等[21] 提出 ABME, 通过不对称的双边估计来合成输入帧之间的中间帧. ABME 首先预测对称双边运动场以获得锚帧 (临时的中间帧). 随后, 从锚帧到输入帧估计不对称双边运动场, 利用这些不对称场对输入帧进行后向翘曲, 从而重建中间帧. 尤其是在快速运动或存在遮挡的情况下, 这种设计思想可以帮助模型更加准确地捕捉复杂运动. 针对视频插帧应用场景对实时性的要求, 马境远等[57] 提出一种多尺度光流预测与融合的实时视频插帧方法, 通过多尺度特征提取和注意力机制增强光流估计, 该方法实现了高质量插帧效果, 并达到实时处理的速率. 为提升视频插帧算法的实时性, 杨华等[58] 提出 SKFEVI 算法, 该算法使用注意力机制进行特征融合, 增强特征表达能力, 只需要进行一次光流估计, 即可得到两个视频帧之间的运动信息. 2022 年, 基于高效的编码器–解码器的网络结构, Kong 等[25] 提出 IFRNet, 通过逐渐细化中间帧特征和双边光流来实现对上下文细节的保留. 此外, IFRNet 在预测可见性掩膜的同时, 还预测中间帧图像的残差, 用于补充合成帧的细节. 同时, 进一步提出一种新颖的面向任务的光流蒸馏损失, 以促进

模型专注于学习有用的教师知识. 针对大运动场景插帧问题, Reda 等[26] 提出 FILM, 其是一个统一的网络, 特点是多尺度特征提取器在所有尺度上共享权重. 值得注意的是, FILM 还使用度量特征映射之间相关差的 Gram 矩阵损失来训练网络, 提高了网络对大运动场景的插帧能力. Ding 等[59] 针对传统双目事件–强度相机系统中由于跨模态视差引发的伪影与失真问题, 提出一种新型的双目事件驱动视频帧插值网络 (Stereo event-based VFI network, SEVFI-Net). 该方法通过特征聚合模块 (Feature aggregation module, FAM) 在特征域内缓解视差, 实现空间对齐, 并利用融合特征精确估计光流和视差, 从而通过基于光流和基于合成的方法生成高质量的中间帧. 基于 IFRNet 和 RAFT 的网络设计思想, Li 等[30] 提出 AMT. 基于构建的 4D 双向相关性矩阵, AMT 使用预测的双边流来检索, 以更新光流和目标帧特征. 得益于采用 IFRNet 逐级细化和 RAFT 构建 4D 双向相关性矩阵的思想, AMT 在建模大运动和处理遮挡的问题上有着较好的表现效果. 随着 Transformer 在计算机视觉中的应用增多, EMA-VFI[31]、VFIFormer[27] 和 BiFormer[32] 尝试将 Transformer 应用于光流估计中. 由于卷积在建模大运动上的局限性, VFIFormer[27] 利用 Transformer 来建模视频帧之间的远程像素相关性, 以获取更加鲁棒的光流. EMA-VFI[31] 通过一个统一的操作显式地提取运动和外观信息, 解决了以往工作中的表示歧义和低效率问题. 基于双向 Transformer, BiFormer[32] 通过全局运动估计和局部运动细化以得到更加准确的光流. 2024 年, 针对视频帧中出现的较大运动问题, 受 Zhao 等[55] 提出的 GMFlow-Net 启发, Liu 等[39] 提出 SGM, 通过稀疏全局匹配算法来优化对大运动插帧场景的处理.

1.4.3 基于前向翘曲的视频插帧方法

对比于后向翘曲来说, 前向翘曲对运动的建模是更加显式的. 2018 年, Niklaus 等[5] 提出一种用于视频插帧的上下文感知合成方法 CS-VFI, 该方法不仅对输入帧进行翘曲, 而且还翘曲它们的逐像素上下文信息, 并使用这些信息来插值生成高质量的中间帧. 该方法的核心在于它采用一种更灵活的合成方法, 可以更好地处理光流估计的不准确性和遮挡问题.

2020 年, Niklaus 等[17] 提出“Softmax splatting”(简称为 SoftSplat) 方法, 该方法特别关注于如何以可微分的方式进行前向翘曲 (即将像素从源位置映射到目标位置), 尤其是当多个源像素映射到同一目标像素时如何处理这种冲突. “Softmax splatting”方法不仅可以提高使用前向翘曲获得的图像质量, 还为后续的基于前向翘曲的视频插帧算法研究提供宝贵的借鉴. 为高效生成中间帧, Hu 等[23] 提出 M2M (Many-to-many) 方法, 不同于传统基于光流的视频插帧方法, 该方法不仅依赖于光流从输入帧向目标插值时刻翘曲像素, 还通过预测每个像素点的多个双向流直接将像素前向翘曲到所需的时间步. 这种方法的关键在于, 每个源像素能够渲染多个目标像素, 每个目标像素可以从更大的视觉上下文区域合成, 这使得插值结果对于遮挡、不连续等光流估计中固有的挑战更为鲁棒. Jin 等[29] 提出一个统一的金字塔循环网络 UPR-Net, 利用轻量级循环模块进行双向流估计和中间流合成. UPR-Net 使用图像金字塔逐级细化的策略, 这使得 UPR-Net 在处理大运动问题上具有较高的优势. 同时, 由于 UPR-Net 各层之间共享参数, 所以使得其参数非常轻量. 针对插帧中由于光流未对齐而导致的模糊和鬼影现象, Wu 等[37] 提出非对称的帧融合模块 PerVFI, 并使用标准化流 (Normalizing flow) 方法生成中间帧, 显著提高了生成帧的质量. 值得注意的是, PerVFI 避免使用重建损失来训练模型, 因为这会导致插帧的模糊. 针对更加复杂的二次曲线运动建模问题, Hu 等[38] 提出 IQ-VFI, 其是一种新的隐式二次视频帧插值框架, 可以更加高效地完成复杂运动插帧问题.

1.5 基于深度学习的各个视频插帧方法对比

如上文所述, 基于深度学习的视频插帧方法主要包括基于相位、核、生成和光流的方法, 各个方法在不同阶段为推动视频插帧技术的发展做出各自的贡献. 基于相位的方法不需显式运动估计, 通过简单的像素级相位调整能快速生成中间帧, 适用于光照变化稳定的场景, 但在大范围运动处理上可能不够精确. 基于核的方法 (如 AdaConv 和 SepConv) 通过学习卷积核知识来捕捉运动信息, 能够有效处理遮挡和动态模糊, 但其对计算资源需求较高. 尤其是在处理较高分辨率视频的情况下, 基于核的方法对计算量的要求是巨大的, 且难以实现实时的插帧性能. 基于生成的方法可以直接预测中间帧, 在场景存在复杂运动和光照变化等情况下优势较为明显, 但可能在未充分学习的动态场景产生不真实的细节. 基于深度学习的光流估计方法能更精准地建模视频运动, 是目前视频插帧领域的研究热点, 也被认为是最具潜力的研究方向. 根据翘曲的方式不同, 基于光流的方法分为前向翘曲和后向翘曲, 前者可以更加显式建模物体运动, 但可能导致空洞问题; 而后者可以避免空洞产生, 但是由于其采用隐式建模的方法, 在任意时刻插帧的应用场景中性能会下降. 此外, 为方便更加深入了解各个方法的优

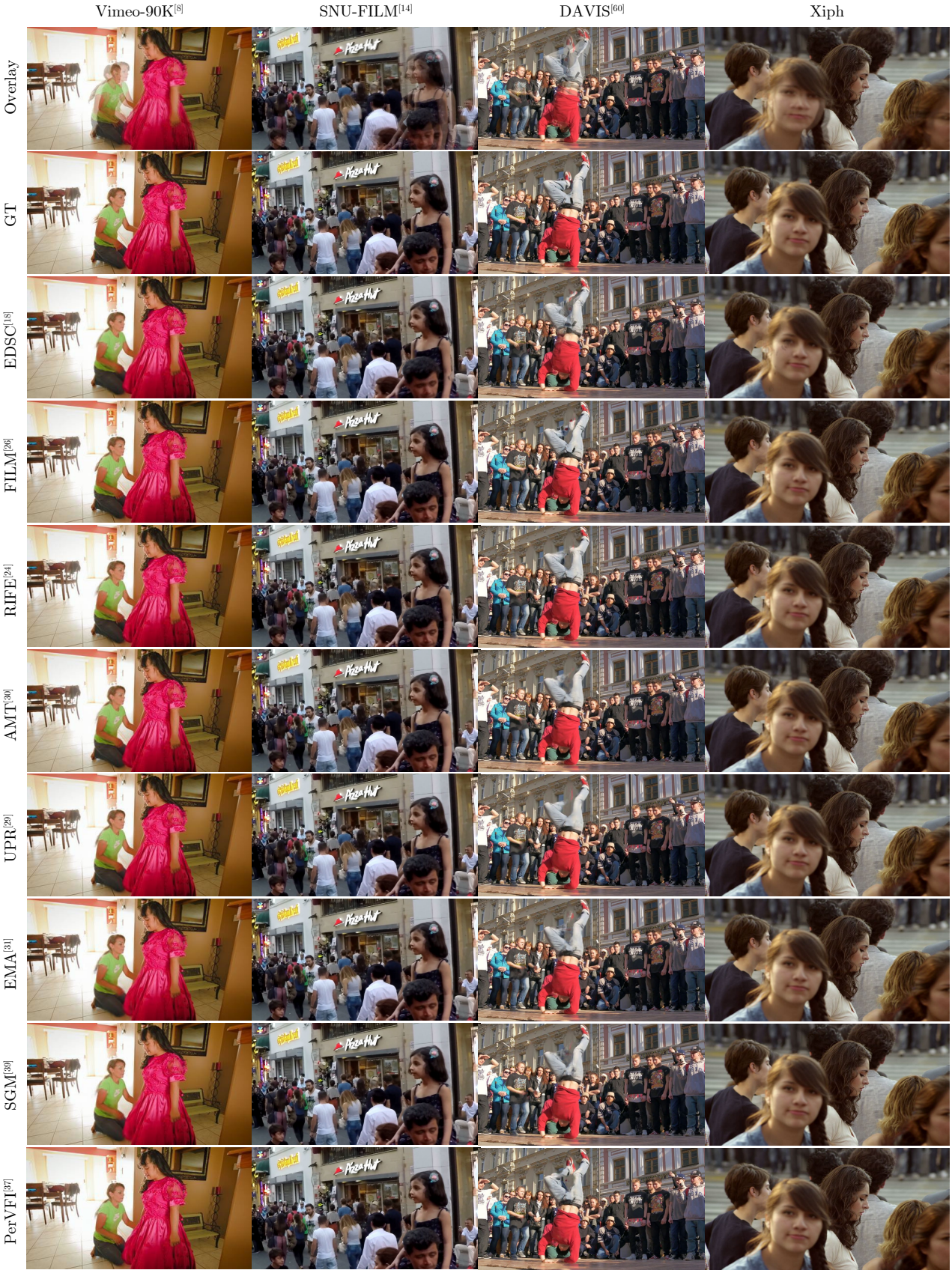


图 2　视频插帧算法在数据集上的可视化结果

Fig. 2　Visualization results of video frame interpolation algorithms on datasets

劣势, 本文给出可视化 (图 2)、算法详细描述 (表 1) 和性能指标对比 (表 2). 在图 2 中, 本文展示部分方法在常用数据集上的可视化结果. 在表 1 中, 按照算法的发表年份, 本文在归类和损失函数等方面给出算法的详细信息. 在表 2 中, 本文给出算法在各个数据集上的性能指标.

表 1 基于深度学习的视频插帧方法对比

Table 1 Comparison of deep-learning based video frame interpolation methods

发表年份	方法	归类	损失函数	训练集	评价指标	训练框架
2017	AdaConv[1]	核	色彩损失, 梯度损失	Flickr	SSIM, IE	PyTorch
	SepConv[2]	核	重构损失	YouTube	PSNR, SSIM, MAE, RMSE	PyTorch
2018	PhaseNet[3]	相位	重构损失, 相位损失	DAVIS	SSIM	Tensorflow
	Super SloMo[4]	光流/后向	重构损失, 感知损失, 翘曲损失, 平滑损失	Adobe240, YouTube240	PSNR, SSIM, IE	PyTorch
	CS-VFI[5]	光流/前向	重构损失, 感知损失, 色彩损失	YouTube	PSNR, SSIM, IE	PyTorch
2019	IM-Net[6]	核	重构损失, 翘曲损失, 相似损失	YouTube	PSNR, SSIM, IE	Caffe
	MEMC-Net[7]	光流/后向	Charbon 损失	Vimeo-90K	PSNR, SSIM, IE	PyTorch
	TOFlow[8]	光流/后向	重构损失	Vimeo-90K	PSNR, SSIM, SSD	PyTorch
	Q-VFI[9]	光流/后向	重构损失, 感知损失	来自互联网	PSNR, SSIM, IE	PyTorch
	DAIN[10]	光流/后向	Charbon 损失	Vimeo-90K	PSNR, SSIM, IE, NIE	PyTorch
2020	FeFlow[11]	核	MMG 损失, 重建损失	Vimeo-90K	PSNR, SSIM, IE	PyTorch
	DSepConv[12]	核	Charbon 损失, 梯度损失	Vimeo-90K	PSNR, SSIM, IE	PyTorch
	AdaCoF[13]	核	重构损失, 失真损失, 感知损失	Vimeo-90K	PSNR, SSIM, IE	PyTorch
	CAIN[14]	生成	重构损失, 感知损失	Vimeo-90K	PSNR, SSIM	PyTorch
	FISR[15]	生成	时间损失, 重构损失	YouTube	PSNR, SSIM	Tensorflow
	BMBC[16]	光流/后向	光度损失, 平滑损失	Vimeo-90K	PSNR, SSIM, IE, NIE	PyTorch
	SoftSplat[17]	光流/前向	色彩损失, 感知损失	Vimeo-90K	PSNR, SSIM, LPIPS	PyTorch
2021	EDSC[18]	核	Charbon 损失, 感知损失	Vimeo-90K	PSNR, SSIM, IE, LPIPS	PyTorch
	CDFI[19]	核	Charbon 损失, 感知损失, 偏移损失	Vimeo-90K	PSNR, SSIM, LPIPS	PyTorch
	XVFI[20]	光流/后向	重构损失, 平滑损失	X-TRAIN	PSNR, SSIM, tOF, EPE	PyTorch
	ABME[21]	光流/后向	Charbon 损失, Census 损失	Vimeo-90K	PSNR, SSIM	PyTorch
2022	VFIT[22]	生成	重构损失	Vimeo-90K	PSNR, SSIM	PyTorch
	M2M[23]	光流/前向	Charbon 损失, Census 损失	Vimeo-90K	PSNR, SSIM	PyTorch
	RIFE[24]	光流/后向	重构损失, 蒸馏损失	Vimeo-90K	PSNR, SSIM, IE	PyTorch
	IFRNet[25]	光流/后向	Charbon 损失, Census 损失, 蒸馏损失, 几何一致性损失	Vimeo-90K	PSNR, SSIM, IE, NIE	PyTorch
	FILM[26]	光流/后向	重构损失, 感知损失, Gram 损失	Vimeo-90K	PSNR, SSIM	Tensorflow
	VFIFormer[27]	光流/后向	重构损失, Census 损失, 蒸馏损失	Vimeo-90K	PSNR, SSIM	PyTorch
2023	FLAVR[28]	生成	重构损失	GoPro, Vimeo-90K	PSNR, SSIM, TCC	PyTorch
	UPR-Net[29]	光流/前向	Charbon 损失, Census 损失	Vimeo-90K	PSNR, SSIM	PyTorch
	AMT[30]	光流/后向	Charbon 损失, Census 损失, 光流损失	Vimeo-90K	PSNR, SSIM	PyTorch
	EMA-VFI[31]	光流/后向	重构损失, 蒸馏损失, 色彩损失, 感知损失	Vimeo-90K	PSNR, SSIM	PyTorch
	BiFormer[32]	光流/后向	Charbon 损失, Census 损失	X-TRAIN	PSNR, SSIM	PyTorch
2024	MSEConv[33]	核	重构损失, 感知损失, 对抗损失	Vimeo-90K	PSNR, SSIM	PyTorch
	LDMVFI[34]	生成	LDM 损失	Vimeo-90K	LPIPS, FloLPIPS, FID	PyTorch
	SwinCS-VFIT[35]	生成	重构损失	Vimeo-90K	PSNR, SSIM	PyTorch
	VFIMamba[36]	生成	拉普拉斯损失, 翘曲损失	X-Train, Vimeo-90K	PSNR, SSIM	PyTorch
	PerVFI[37]	光流/前向	负对数似然损失, 感知损失	Vimeo-90K	PSNR, SSIM, LPIPS, FloLPIPS, VFIPS	PyTorch
	IQ-VFI[38]	光流/前向	重构损失, 蒸馏损失	Vimeo-90K	PSNR, SSIM	PyTorch
	SGM[39]	光流/后向	重构损失, 翘曲损失	X-Train, Vimeo-90K	PSNR, SSIM	PyTorch

表 2 基于深度学习的视频插帧方法性能对比 (评价指标: PSNR ↑ /SSIM ↑ /LPIPS ↓)

Table 2 Performance comparison of deep-learning based video frame interpolation methods (Evaluation metrics: PSNR ↑ /SSIM ↑ /LPIPS ↓)

发表年份	方法	Vimeo-90K	UCF101	X-TEST	Xiph	DAVIS	SNU-FILM			
							Easy	Medium	Hard	Extreme
2017	AdaConv[1]	32.33/0.957/–	—	—	—	—	—	—	—	—
	SepConv[2]	33.45/0.967/0.019	33.02/0.935/0.024	24.34/0.742/–	32.61/0.880/–	26.21/0.857/–	39.68/0.990/–	35.07/0.976/–	29.39/0.926/–	34.32/0.845/–
2018	Super SloMo[4]	32.90/0.957/–	33.14/0.938/–	—	—	25.76/0.850/–	37.28/0.986/–	33.80/0.973/–	28.98/0.925/–	24.15/0.845/–
2019	MEMC-Net[7]	34.02/0.970/0.027	34.95/0.968/0.030	—	—	—	—	—	—	—
	IM-Net[6]	33.50/0.947/–	—	—	—	—	—	—	—	—
	TOFlow[8]	33.53/0.967/0.027	34.58/0.967/0.027	—	—	—	39.08/0.989/–	34.39/0.974/–	28.44/0.918/–	23.39/0.831/–
	Q-VFI[9]	35.15/0.971/–	32.54/0.948/–	—	—	27.73/0.894/–	—	—	—	—
	DAIN[10]	34.71/0.976/0.022	34.99/0.968/0.028	26.78/0.807/–	—	26.12/0.870/–	39.73/0.990/–	35.46/0.978/–	30.17/0.934/–	25.09/0.858/–
2020	FeFlow[11]	35.28/0.976/–	—	24.00/0.756/–	—	—	—	—	—	—
	DSepConv[12]	34.73/0.974/0.028	35.08/0.969/0.030	—	—	—	—	—	—	—
	AdaCoF[13]	35.40/0.971/0.031	35.06/0.974/0.033	24.13/0.734/–	32.72/0.881/–	27.07/0.874/–	39.80/0.990/0.019	35.05/0.975/0.036	29.46/0.924/0.075	24.30/0.844/0.148
	CAIN[14]	34.65/0.973/–	34.91/0.969/–	24.50/0.752/–	24.50/0.752/–	26.46/0.856/–	39.78/0.990/–	35.49/0.977/–	29.86/0.929/–	24.69/0.850/–
	FISR[15]	—	—	—	—	—	—	—	—	—
	BMBC[16]	35.01/0.976/–	32.61/0.955/0.032	22.86/0.727/–	31.27/0.880/–	26.42/0.868/–	39.89/0.990/0.018	35.31/0.977/0.034	29.32/0.927/0.075	23.92/0.843/0.152
	SoftSplat[17]	35.48/0.964/0.013	35.10/0.948/0.022	25.48/0.725/–	—	27.42/0.878/–	—	—	—	—
2021	EDSC[18]	34.84/0.975/0.026	35.13/0.968/0.029	—	—	24.54/0.768/0.205	—	—	—	—
	CDFI[19]	35.17/0.964/0.010	35.21/0.950/0.015	24.49/0.742/–	33.01/0.872/–	—	40.11/0.990/0.013	35.50/0.978/0.024	29.74/0.928/0.056	24.54/0.847/0.121
	XVFI[20]	35.07/0.968/–	32.65/0.968/0.033	30.12/0.870/–	34.06/0.895/–	—	39.55/0.989/0.020	35.06/0.976/0.037	29.51/0.927/0.075	24.43/0.848/0.143
	ABME[21]	36.18/0.981/–	32.05/0.967/0.058	30.16/0.879/–	33.81/0.903/–	—	39.69/0.990/0.022	35.28/0.977/0.042	29.64/0.929/0.092	24.54/0.853/0.182
2022	RIFE[24]	35.61/0.978/0.020	35.28/0.969/–	24.67/0.797/–	—	25.89/0.803/0.134	40.06/0.991/–	35.75/0.979/–	30.10/0.933/–	24.84/0.853/–
	VFIT[22]	36.96/0.978/–	33.44/0.971/–	—	—	28.09/0.888/–	—	—	—	—
	IFRNet[25]	36.20/0.981/–	35.42/0.970/0.031	30.46/–/–	—	—	40.10/0.991/0.017	36.12/0.980/0.029	30.63/0.937/0.058	25.27/0.861/0.128
	FILM[26]	35.87/0.968/–	35.16/0.949/–	—	—	—	—	—	—	—
	M2M[23]	35.40/0.978/–	35.17/0.970/–	30.81/0.912/–	34.46/0.925/–	—	39.66/0.991/–	35.74/0.980/–	30.32/0.936/–	25.07/0.860/–
	VFIFormer[27]	36.50/0.982/0.021	35.43/0.970/0.034	24.58/0.805/–	33.69/0.925/–	—	40.13/0.991/0.018	36.09/0.980/0.033	30.67/0.938/0.069	25.43/0.864/0.146
2023	FLAVR[28]	36.25/0.975/–	33.31/0.971/–	—	—	27.43/0.874/–	—	—	—	—
	AMT[30]	36.53/0.982/0.021	35.45/0.970/–	—	—	—	39.88/0.991/–	36.12/0.981/–	30.78/0.939/–	25.43/0.865/–
	EMA-VFI[31]	36.64/0.982/0.026	35.48/0.970/–	31.46/–/–	—	37.61/0.846/0.203	39.98/0.991/–	36.09/0.980/–	30.94/0.939/–	25.69/0.866/–
	BiFormer[32]	—	—	31.32/0.921/–	34.48/0.927/–	—	—	—	—	—
	UPR-Net[29]	36.42/0.982/–	35.47/0.970/–	30.50/0.905/–	—	—	40.44/0.991/–	36.29/0.980/–	30.86/0.938/–	25.63/0.864/–
2024	LDMVFI[34]	—	32.16/0.964/0.026	—	—	—	38.89/0.988/0.013	33.97/0.971/0.027	28.14/0.911/0.068	23.34/0.827/0.139
	SGM[39]	—	—	29.91/0.897/–	29.25/0.818/–	—	40.15/0.991/–	36.05/0.980/–	28.88/0.922/–	23.62/0.838/–
	PerVFI[37]	33.89/0.953/0.018	—	—	—	26.23/0.808/0.114	—	—	—	—
	IQ-VFI[38]	36.60/0.982/–	35.48/0.970/–	—	—	—	40.24/0.991/–	36.24/0.980/–	30.83/0.938/–	25.45/0.863/–
	SwinCS-VFIT[35]	37.13/0.978/–	33.36/0.971/–	—	—	28.28/0.891/–	—	—	—	—
	VFIMamba[36]	36.64/0.982/–	35.45/0.970/–	32.15/0.925/–	34.62/0.906/–	—	40.51/0.991/–	36.40/0.981/–	30.99/0.940/–	25.79/0.868/–
	MSEConv[33]	—	35.10/0.966/–	—	—	—	—	—	—	—

2 视频插帧常用数据集

如表 3 所示, 为更好地研究基于深度学习的视频插帧, 构建大规模的数据集是必然要求, 目前有多个视频插帧的公共数据集供研究人员使用.

1) Xiph 数据集: 该数据集提出的目的是测试图像压缩算法的性能. 在 X4K1000FPS 数据集出现之前, 由于没有专门为评估视频插帧性能而提出的高分辨率 (4 K) 视频插帧数据集, Xiph 数据集被用来进行视频插帧算法在高分辨率场景下的评估. 在使用该数据集评估视频插帧算法性能时, 由于算法所需显存大小等问题, 常通过改变大小或中心裁切将其处理成 2 K 的版本.

2) Middlebury 数据集[61]: Middlebury 是在视频插帧技术早期广泛使用的数据集, 近几年的视频插帧算法较少使用该数据集进行评估. 该数据集的图像分辨率约为 640 × 480 像素, 它通常仅用于评估 8 个序列的 VFI 方法.

3) UCF101 数据集[62]: UCF101 数据集最初是为人类动作识别而收集的数据集, 包含各种人类行为的 13 320 个视频. 该数据集常用于评估视频插帧方法, 其作为视频插帧任务测试集包含 379 个三元组, 每张图像大小为 256 × 256 像素, 运动幅度和复杂度高于 Vimeo-90K 数据集, 但是由于其分辨率较低, 和现实应用场景的分辨率有较大区别, 不能较好反映视频插帧算法在实际场景下的性能.

4) DAVIS 数据集[60]: DAVIS 数据集是 2016 年提出的, 由 90 个高质量视频序列组成, 用于视频中对象的分割任务. 由于该数据集包含较大、较为复杂的运动, 被近期提出的插帧模型较为广泛地用于评估模型性能. 官方提供的 DAVIS 数据集包含两个版本 (480 P 和全分辨率). 480 P 版本的各视频分辨率不一, 但各视频帧的宽度统一为 480 像素, 常用于显存占用较大模型 (如基于扩散模型的视频插帧技术) 的评估. 同样地, 全分辨率版本的各视频分辨率不一, 最高分辨率为 4 096 × 2 160 像素, 较低分辨率 (占比较多) 为 1 920 × 1 080 像素. 一些方法 (如 PerVFI) 将高于 1 080 P 分辨率的视频处理成 1 080 P, 以评估视频插帧方法的性能. 由于较大的分辨率、运动距离和运动复杂度, 视频插帧算法在该数据集上存在较大的挑战, 在该数据集上的性能评估能较好地反映视频插帧算法的性能.

5) GOPRO 数据集[63]: GOPRO 数据集由手持式 GOPRO4 Hero Black 相机进行采集, 采集的总视频数目为 33 个, 被切分为 3 214 对模糊和清晰图像组成, 分辨率为 1 280 × 720 像素. 该数据集包含在室内室外场景的多种复杂运动, 目前视频插帧算法的性能在该数据集上具有挑战性.

6) Adobe240 数据集[64]: Adobe240 数据集最初用于视频去模糊, 是 VFI 的另一个广泛使用的数据集. 它由分辨率为 1 280 × 720 的高帧率视频 (240 FPS) 组成, 但视频仅来自 118 个片段. 该数据集是视频插帧技术早期使用的评估数据集, 测试视频分辨率高于 Middlebury, 近期提出的视频插帧算法并未在该数据集上进行评测.

7) Vimeo-90K 数据集[8]: Vimeo-90K 数据集是在 Vimeo 视频共享平台下载, 专门为视频处理任务设计的数据集, 被广泛用于视频超分、视频插帧等领域. 该数据集是用于训练和评估视频插帧模型最常用的数据集. 它包含 73 171 个三元组数据, 每张图像大小为 448 × 256 像素, 从现实生活视频剪辑中的 4 278 个视频中提取, 包含大量的视频中存在的场景. 在视频插帧的领域研究中, 常使用 Vimeo-90K 数据集作为训练集, 在该数据集上训练的模型

表 3 基于深度学习的视频插帧技术使用的数据集

Table 3 Datasets used for deep-learning based video frame interpolation technology

数据集	发布年份	视频数目	分辨率 (像素)	常用评价指标
Xiph	1994	8	4 096 × 2 160	PSNR, SSIM, IPIPS
Middlebury[61]	2011	24	640 × 480	IE
UCF101[62]	2012	13 320	256 × 256	PSNR, SSIM, IPIPS
DAVIS[60]	2016	90	4 096 × 2 160	PSNR, SSIM, IPIPS
GOPRO[63]	2017	33	1 280 × 720	PSNR, SSIM, IE
Adobe240[64]	2017	71	1 280 × 720	PSNR, SSIM, IE
Vimeo-90K[8]	2019	4 278	448 × 256	PSNR, SSIM, IPIPS
HD[7]	2019	7	1 280 × 720	PSNR
SNU-FILM[14]	2020	31	1 280 × 720	PSNR, SSIM, IPIPS
X4K1000FPS (X-TEST)[20]	2021	15	4 096 × 2 160	PSNR, SSIM, IPIPS
SportsSloMo[65]	2024	8 498	1 280 × 720	PSNR, SSIM, IE

在其他数据集上进行测试. Vimeo-90K 的训练集由于其包含广泛的场景、数据量较大的特点, 依旧适合于进行插帧模型的训练. 近年来, 随着视频插帧技术的进步, 由于 Vimeo-90K 数据集分辨率较小、整体数据集包含的运动幅度较小, 各方法在 Vimeo-90K 数据集上的测试表现相差较小, 已经不能满足视频插帧技术发展的评测需求.

8) HD 数据集[7]: HD 数据集收集自 Xiph 网站的 7 个高分辨率视频, 该数据集中的运动较大. 但随着更高质量的测试数据集的提出, HD 数据集较少用于评估近年来提出的视频插帧算法的性能.

9) SNU-FILM 数据集[14]: SNU-FILM 数据集是一个广泛使用的视频插帧评估基准, 该数据集由 31 个 240 FPS 的视频序列组成, 包含来自 GOPRO 数据集的 11 个视频和来自 YouTube 的 20 个视频. 该数据集包含 1 240 个三元组数据, 每张图像具有 1 280 × 720 像素的分辨率. 并根据运动幅度分为 Easy (简单)、Medium (中等)、Hard (困难) 和 Extreme (极其困难) 四个不同的部分. 视频插帧算法的性能对比在 Easy 部分和 Medium 部分的测试中相差较小, 但是在 Hard 部分和 Extreme 部分上相差较大, 且对视频插帧方法性能的评估仍发挥着重要的作用.

10) X4K1000FPS 数据集[20]: X4K1000FPS 数据集是 2021 年发布的数据集, 该数据集拥有 4 K 的图像分辨率, 视频帧之间物体的运动有着较高的位移. 在视频插帧领域, 常使用 X4K1000FPS 的测试集部分 X-TEST 作为测试集, 来测试视频插帧模型在高分辨率、大位移场景中的表现, 目前的视频插帧算法还没有在该数据集上有较好的表现, 适合于未来评测视频插帧算法性能.

11) SportsSloMo 数据集[65]: 于 2024 年发布的大型数据集, 采集来自 YouTube 中的各种比赛运动场景, 是以人为中心的视频插帧数据集. 因为人体是高度可变形的, 并且在运动中遮挡是常见的, 这对视频插帧算法提出挑战. 该数据集满足现有插帧算法的发展需求, 未来视频插帧模型的训练可能同时采用 SportsSloMo 和 Vimeo-90K 数据集, 可以有效解决 Vimeo-90K 数据集复杂运动场景较少的问题, 提升视频插帧模型的性能.

综上所述, Middlebury、Adobe240 和 HD 数据集作为早期用于评测视频插帧算法性能的数据集, 为视频插帧的发展做出重要的贡献. 但是随着视频插帧技术的发展, 由于较小分辨率、较低的视频质量和较为单一的评估场景, 这些数据集较少被用于评估视频插帧算法的性能.Vimeo-90K 是近年来较为广泛使用的视频插帧算法训练集, 但是由于其低分辨率的特点, 各方法在其测试集上性能趋于饱和, 未来不太适合进行视频插帧的算法评估. 与 Vimeo-90K 测试集类似, UCF101 因为其分辨率低和场景单一的缺点, 其不太适用于未来视频插帧算法的评估. GOPRO、SNU-FILM 和 SportsSloMo 数据集是面向普通分辨率场景的评估数据集, 其多样的场景和复杂的运动对视频插帧算法具有较大的挑战性. Xiph、X4K1000FPS 和 DAVIS 数据集由于其高分辨率和存在复杂运动场景等特点, 与现实应用有着较好的吻合性, 在这些数据集上的评估结果具有较高的可比性. 同时, Kiefhaber 等[66] 指出当前视频插帧技术的评估存在诸多问题, 包括测试数据集繁多、误差指标计算方式不一致, 难以在不同方法间进行公平对比. 此外, 许多测试集的提出服务于论文提出方法的效果验证, 缺乏专门基准测试的深入评估, 还因违背线性假设导致无法在无先验知识的情况下解决这些问题. 为此, Kiefhaber 等[66] 提出一个新的基准测试框架, 通过一致的误差指标计算、基于线性假设的精心设计测试集 (利用合成数据)、插帧质量与像素属性分析以及计算效率评估, 为社区提供一个系统化的评估工具.

3 思考与展望

深度学习技术在视频插帧领域展现出广泛的应用潜力, 并已成为当前研究的热点. 许多研究通过在深度网络架构的构建、损失函数的设计以及训练策略的优化等方面做出富有成效的探索, 对视频插帧方法的发展做出了重要贡献. 展望未来, 以下几个方面仍然充满挑战与机遇:

1) 运动的复杂性: 对于视频插帧算法的运动估计, 尽管其在普通的场景下表现良好, 但是由于其假设条件过于理想化, 难以适应复杂的现实运动场景. 以下列举多个复杂的运动场景:

a) 非线性运动: 基于光流的视频插帧算法通常假设物体运动在短时间内是线性的, 但是在实际的插帧任务上, 许多任务是非线性的. 例如物体的运动速度在多个方向存在加速度, 从而导致物体沿着曲线轨迹移动, 甚至在短时间内发生不规则抖动. 这种非线性运动超出线性光流模型的能力范围, 会导致插帧结果的连续性受到严重影响.

b) 遮挡问题: 在实际场景中, 物体之间相对运动并发生遮挡是常见的情况. 基于光流的视频插帧任务难以准确捕捉被遮挡物体的运动信息, 而遮挡区域可能导致光流计算的不确定性. 目前的插帧算法在面对包含遮挡问题的插帧场景问题时, 大多都会表现出较为明显的伪影, 严重影响了插帧质量.

c) 纹理相似的场景: 基于光流的视频插帧任务在估计光流时严重依赖图像纹理的局部特征, 在纹

理重复出现的区域 (例如墙壁、斑马线等), 运动估计的可靠性将会被降低, 容易出现翘曲后物体运动的漂移或不合理的运动.

d) 物体畸变: 物体畸变包含多种情况, 通常由两方面的因素导致, 即拍摄设备和物体运动. 对于拍摄设备来说, 当使用的设备镜头拍摄的画面本就存在畸变 (如鱼眼镜头), 这会导致拍摄的画面被扭曲, 同一物体在前帧和后帧中发生不同程度的扭曲, 进而导致物体特征不匹配, 从而影响了光流的准确性. 由于物体运动导致畸变的情况在现实中广泛存在, 例如物体面向镜头转动和移动等. 物体面向镜头转动会导致边缘纹理被扭曲, 导致同一物体在前帧和后帧中纹理并不匹配, 从而影响了光流匹配的准确性. 同时, 当物体面向或者远离镜头移动时, 这会造成物体纹理的放大与缩小, 从而影响了前后帧纹理的一致性, 导致插帧结果质量的降低.

2) 特殊场景插帧: 在视频插帧的研究中, 特殊场景的插帧问题往往因其复杂性和较小的研究关注度而被忽视. 然而, 这些场景具有重要的实际意义, 因为它们在多种应用场景中经常出现, 例如电影制作、视频编辑、游戏动画和虚拟现实. 以下列举多个特殊插帧场景:

a) 多图层运动: 多图层运动是指在同一视频帧中, 不同图层或者深度平面上的物体以各自不同的速度和方向运动. 如影子在地面上的移动、玻璃反射图像的运动等, 在这些运动场景中, 同一像素同时表示多个运动各异图层. 而基于光流的视频插帧算法通常对一个场景估计一个光流, 这会导致不符合估计光流运动的图层被破坏. 基于光流的视频插帧算法难以同时捕捉多个图层 (如前景和背景) 的独立运动信息, 容易将不同图层的运动混淆, 从而影响了插帧结果的质量.

b) 物体发生突变的场景: 物体发生突变的场景是视频插帧技术中极具挑战性的问题之一, 其复杂性源于运动模式的突然改变和光照条件的剧烈变化. 这类场景在现实中广泛存在, 如舞台灯光的突然切换、电影字幕的替换以及灯光瞬间熄灭或亮起等. 这些突变不仅改变视频帧的整体亮度和纹理分布, 还可能导致物体轮廓的显著变化, 对插帧算法的运动估计和纹理生成能力提出极高要求.

c) 前后帧画质不对齐: 前后帧画质不对齐也是插帧技术中常见但难以克服的障碍. 这种情况通常出现在编码压缩率变化、动态分辨率调整或噪声干扰显著的场景中. 不同画质的前后帧对运动估计造成巨大的挑战, 高质量帧提供的纹理信息可能在低质量帧中完全丢失, 导致光流模型生成的运动矢量充满误差, 插帧结果因此出现伪影或分辨率不统一的问题. 即便在高性能模型中, 针对不同质量帧进行插帧时, 结果的稳定性和一致性也常常难以保证.

3) 实时处理能力: 在视频插帧领域, 实时处理能力的重要性日益凸显, 尤其是在移动端和嵌入式设备等低算力环境中应用时, 这一需求更为迫切. 随着深度学习模型的广泛应用, 插帧算法的性能和插帧质量得到显著提升, 但随之而来的计算复杂度和参数规模的增长也带来新的挑战. 在高性能计算设备上运行的插帧算法往往难以迁移至低功耗设备, 限制其在视频直播、移动视频编辑以及虚拟现实等场景中的实时应用.

4) 更高分辨率视频的插帧: 随着 4 K、8 K 等高分辨率视频的普及, 视频插帧技术面临着新的挑战. 在高分辨率视频中, 画面包含大量的细节信息, 物体运动的尺度也显著增加, 这对插帧算法的运动估计、细节保留和计算效率提出更高的要求. 如何在不牺牲细节和插帧质量的前提下高效处理超高分辨率视频, 已成为当前研究的热点问题.

5) 基于大模型技术的视频插帧: 随着以 Sora 为首的视频生成大模型迅速走红, 人们逐渐认识到这类模型在多个领域的广泛应用潜力. 大模型技术的出现深刻地改变视频处理领域的许多传统方法, 同时也对视频插帧领域产生一定影响, 大模型技术将在如下几个方面对视频插帧领域产生深刻影响:

a) 更强的表示能力与特征学习能力: 大模型的最大特点之一是其庞大的参数量和深度结构, 这使得它们能够更好地捕捉视频数据中的高层次特征和复杂的时空依赖关系. 传统的视频插帧方法通常依赖于端到端学习的浅层特征和简单的运动估计来生成中间帧, 然而这种方法在处理复杂场景或高动态范围时容易产生失真和伪影. 大模型通过大量的数据训练, 可以自动学习视频中的丰富时空信息, 不仅能精确捕捉图像中的局部细节, 还能深入理解视频中的全局动态, 提升了插帧的质量.

b) 增强的感知优化能力: 传统的视频插帧方法通常依赖于像 PSNR、SSIM 等客观指标进行优化, 这些指标虽然能衡量图像质量, 但与人类的感知质量存在较大差距, 特别是在复杂视频场景下, 简单的像素级优化往往不能充分提升感知效果. 大模型技术, 尤其是生成模型 (扩散模型), 通过感知优化目标的引入, 使得模型不仅关注像素精度, 还能优化视频的整体感知质量.

c) 高效的时空建模: 视频插帧涉及的不仅仅是空间图像生成, 还需要准确建模时间维度上的动态变化. 传统方法通常依赖光流估计或简单的运动插值来生成中间帧, 这在面对快速运动、复杂变形或场景遮挡时可能失效. 大模型的时空建模能力能够

精确捕捉视频帧之间的运动模式, 并推断出如何在时间维度上插入中间帧. 这种能力使得大模型在处理高动态视频内容时, 能够有效地预测运动轨迹, 并生成与前后帧之间自然连接的中间帧. 尤其是在面对复杂的动态纹理、快速运动或大范围遮挡时, 大模型能够通过全局上下文和长时间序列学习, 显著提高插值结果的准确性和视觉质量.

视频插帧技术在深度学习的推动下取得显著进展, 但仍面临复杂运动场景、特殊场景插帧、实时处理能力和高分辨率视频处理等挑战. 复杂运动如非线性运动和遮挡问题影响了光流方法的精度, 而特殊场景如多图层运动和突变情况也增加了插帧难度. 实时处理在低算力设备上的应用受限, 高分辨率视频则要求更高的计算效率和细节保留. 大模型技术的兴起为视频插帧带来新的机遇, 它通过更强的表示能力、感知优化能力和高效的时空建模, 显著提高了插帧结果的质量. 未来, 优化算法效率、提高适应性、结合深度学习与硬件加速, 并探索大模型的潜力, 将是推动视频插帧技术发展的关键问题.

References

1 Niklaus S, Mai L, Liu F. Video frame interpolation via adaptive convolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 2270−2279

2 Niklaus S, Mai L, Liu F. Video frame interpolation via adaptive separable convolution. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 261−270

3 Meyer S, Djelouah A, McWilliams B, Sorkine-Hornung A, Gross M, Schroers C. PhaseNet for video frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 498−507

4 Jiang H Z, Sun D Q, Jampani V, Yang M H, Learned-Miller E, Kautz J. Super SloMo: High quality estimation of multiple intermediate frames for video interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 9000−9008

5 Niklaus S, Liu F. Context-aware synthesis for video frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 1701−1710

6 Peleg T, Szekely P, Sabo D, Sendik O. IM-Net for high resolution video frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 2393−2402

7 Bao W B, Lai W S, Zhang X Y, Gao Z Y, Yang M H. MEMC-Net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, **43**(3): 933−948

8 Xue T F, Chen B A, Wu J J, Wei D L, Freeman W T. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 2019, **127**(8): 1106−1125

9 Xu X Y, Li S Y, Sun W X, Yin Q, Yang M H. Quadratic video interpolation. In: Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS). Vancouver, Canada: MIT Press, 2019. 1645−1654

10 Bao W B, Lai W S, Ma C, Zhang X Y, Gao Z Y, Yang M H. Depth-aware video frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 3698−3707

11 Gui S R, Wang C Y, Chen Q H, Tao D C. FeatureFlow: Robust video interpolation via structure-to-texture generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 14001−14010

12 Cheng X H, Chen Z Z. Video frame interpolation via deformable separable convolution. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020. 10607−10614

13 Lee H, Kim T, Chung T Y, Pak D, Ban Y, Lee S. AdaCoF: Adaptive collaboration of flows for video frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 5315−5324

14 Choi M, Kim H, Han B, Xu N, Lee K M. Channel attention is all you need for video frame interpolation. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020. 10663−10671

15 Kim S Y, Oh J, Kim M. FISR: Deep joint frame interpolation and super-resolution with a multi-scale temporal loss. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020. 11278−11286

16 Park J, Ko K, Lee C, Kim C S. BMBC: Bilateral motion estimation with bilateral cost volume for video interpolation. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020. 109−125

17 Niklaus S, Liu F. Softmax splatting for video frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 5436−5445

18 Cheng X H, Chen Z Z. Multiple video frame interpolation via enhanced deformable separable convolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, **44**(10): 7029−7045

19 Ding T Y, Liang L M, Zhu Z H, Zharkov I. CDFI: Compression-driven network design for frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 7997−8007

20 Sim H, Oh J, Kim M. XVFI: Extreme video frame interpolation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 14469−14478

21 Park J, Lee C, Kim C S. Asymmetric bilateral motion estimation for video frame interpolation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 14519−14528

22 Shi Z H, Xu X Y, Liu X H, Chen J, Yang M H. Video frame interpolation transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 17461−17470

23 Hu P, Niklaus S, Sclaroff S, Saenko K. Many-to-many splatting for efficient video frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 3543−3552

24 Huang Z W, Zhang T Y, Heng W, Shi B X, Zhou S C. Real-time intermediate flow estimation for video frame interpolation. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 624−642

25 Kong L T, Jiang B Y, Luo D H, Chu W Q, Huang X M, Tai Y, et al. IFRNET: Intermediate feature refine network for efficient frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 1959−1968

26 Reda F, Kontkanen J, Tabellion E, Sun D Q, Pantofaru C, Curless B. FILM: Frame interpolation for large motion. In: Proceed-

ings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 250−266

27 Lu L Y, Wu R Z, Lin H J, Lu J B, Jia J Y. Video frame interpolation with transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 3522−3532

28 Kalluri T, Pathak D, Chandraker M, Tran D. FLAVR: Flow-agnostic video representations for fast frame interpolation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2023. 2070−2081

29 Jin X, Wu L H, Chen J, Chen Y X, Koo J, Hahm C H. A unified pyramid recurrent network for video frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 1578−1587

30 Li Z, Zhu Z L, Han L H, Hou Q B, Guo C L, Cheng M M. AMT: All-pairs multi-field transforms for efficient frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 9801−9810

31 Zhang G Z, Zhu Y H, Wang H N, Chen Y X, Wu G S, Wang L M. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 5682−5692

32 Park J, Kim J, Kim C S. BiFormer: Learning bilateral motion estimation via bilateral transformer for 4K video frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 1568−1577

33 Ding X L, Huang P, Zhang D Y, Liang W, Li F, Yang G B, et al. MSEConv: A unified warping framework for video frame interpolation. *ACM Transactions on Asian and Low-resource Language Information Processing*, DOI: 10.1145/3648364

34 Danier D, Zhang F, Bull D. LDMVFI: Video frame interpolation with latent diffusion models. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2024. 1472−1480

35 Shi Chang-Tong, Shan Hong-Tao, Zheng Guang-Yuan, Zhang Yu-Jin, Liu Huai-Yuan, Zong Zhi-Hao. Video frame interpolation method based on improved visual Transformer. *Application Research of Computers*, 2024, **41**(4): 1252−1257
(石昌通, 单鸿涛, 郑光远, 张玉金, 刘怀远, 宗智浩. 改进视觉Transformer的视频插帧方法. 计算机应用研究, 2024, **41**(4): 1252−1257)

36 Zhang G Z, Liu C X, Cui Y T, Zhao X T, Ma K, Wang L M. VFIMamba: Video frame interpolation with state space models. In: Proceedings of the 38th Conference on Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2024. 1645−1654

37 Wu G Y, Tao X, Li C L, Wang W Y, Liu X H, Zheng Q Q. Perception-oriented video frame interpolation via asymmetric blending. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 2753−2762

38 Hu M S, Jiang K, Zhong Z H, Wang Z, Zheng Y Q. IQ-VFI: Implicit quadratic motion estimation for video frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 6410−6419

39 Liu C X, Zhang G Z, Zhao R, Wang L M. Sparse global matching for video frame interpolation with large motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 19125−19134

40 Parihar A S, Varshney D, Pandya K, Aggarwal A. A comprehensive survey on video frame interpolation techniques. *The Visual Computer*, 2022, **38**(1): 295−319

41 Dong J, Ota K, Dong M X. Video frame interpolation: A comprehensive survey. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023, **19**(2s): Article No. 78

42 Meyer S, Wang O, Zimmer H, Grosse M, Sorkine-Hornung A. Phase-based frame interpolation for video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 1410−1418

43 Prashnani E, Noorkami M, Vaquero D, Sen P. A phase-based approach for animating images using video examples. *Computer Graphics Forum*, 2017, **36**(6): 303−311

44 Wadhwa N, Rubinstein M, Durand F, Freeman W T. Phase-based video motion processing. *ACM Transactions on Graphics (TOG)*, 2013, **32**(4): Article No. 80

45 Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 4489−4497

46 Lin Chuan-Jian, Deng Wei, Tong Tong, Gao Qin-Quan. Blurred video frame interpolation method based on deep voxel flow. *Journal of Computer Applications*, 2020, **40**(3): 819−824
(林传健, 邓炜, 童同, 高钦泉. 基于深度体素流的模糊视频插帧方法. 计算机应用, 2020, **40**(3): 819−824)

47 Cho H, Kim T, Jeong Y, Yoon K J. TTA-EVF: Test-time adaptation for event-based video frame interpolation via reliable pixel and sample estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 25701−25711

48 Dosovitskiy A, Fischer P, Ilg E, Häusser P, Hazirbas C, Golkov V, et al. FlowNet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 2758−2766

49 Ilg E, Mayer N, Saikia T, Keuper M, Dosovitskiy A, Brox T. Flownet 2.0: Evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 2462−2470

50 Ranjan A, Black M J. Optical flow estimation using a spatial pyramid network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 2720−2729

51 Sun D Q, Yang X D, Liu M Y, Kautz J. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 8934−8943

52 Hui T W, Tang X O, Loy C C. LiteFlowNet: A lightweight convolutional neural network for optical flow estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 8981−8989

53 Yang G S, Ramanan D. Volumetric correspondence networks for optical flow. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2019. 794−805

54 Teed Z, Deng J. RAFT: Recurrent all-pairs field transforms for optical flow. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020. 402−419

55 Zhao S Y, Zhao L, Zhang Z X, Zhou E Y, Metaxas D. Global matching with overlapping attention for optical flow estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 17592−17601

56 Zhang Qian, Jiang Feng. Video interpolation based on deep learing. *Intelligent Computer and Applications*, 2019, **9**(4): 252−257
(张倩, 姜峰. 基于深度学习的视频插帧算法. 智能计算机与应用, 2019, **9**(4): 252−257)

57 Ma Jing-Yuan, Wang Chuan-Ming. Real-time video frame interpolation based on multi-scale optical prediction and fusion. *Journal of Chinese Computer Systems*, 2021, **42**(12): 2567−2571
(马境远, 王川铭. 一种多尺度光流预测与融合的实时视频插帧方法. 小型微型计算机系统, 2021, **42**(12): 2567−2571)

58 Yang Hua, Wang Jiao, Zhang Wei-Jun, Wu Jie-Hong, Gao Li-Jun. Lightweight video frame interpolation algorithm based on optical flow estimation. *Journal of Shenyang Aerospace ace University*, 2022, **39**(6): 57−64
(杨华, 王姣, 张维君, 吴杰宏, 高利军. 基于光流估计的轻量级视频插帧算法. 沈阳航空航天大学学报, 2022, **39**(6): 57−64)

59 Ding C, Lin M Y, Zhang H J, Liu J Z, Yu L. Video frame interpolation with stereo event and intensity cameras. *IEEE Transactions on Multimedia*, 2024, **26**: 9187−9202

60 Perazzi F, Pont-Tuset J, McWilliams B, van Gool L, Gross M, Sorkine-Hornung A. A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 724−732

61 Baker S, Scharstein D, Lewis J P, Roth S, Black M J, Szeliski R. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 2011, **92**(1): 1−31

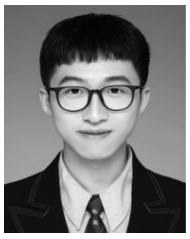
62 Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv: 1212.0402, 2012.

63 Nah S, Kim T H, Lee K M. Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 257−265

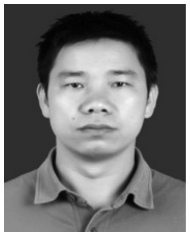
64 Su S C, Delbracio M, Wang J, Sapiro G, Heidrich W, Wang O. Deep video deblurring for hand-held cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 237−246

65 Chen J B, Jiang H Z. SportsSloMo: A new benchmark and baselines for human-centric video frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2024. 6475−6486

66 Kiefhaber S, Niklaus S, Liu F, Schaub-Meyer S. Benchmarking video frame interpolation. arXiv preprint arXiv: 2403.17128, 2024.



吴晨阳 南开大学计算机学院博士研究生. 主要研究方向为深度学习和视频插帧.
E-mail: wucy0519@gmail.com
(**WU Chen-Yang** Ph.D. candidate at the College of Computer Science, Nankai University. His research interest covers deep learning and video frame interpolation.)



张 勇 重庆长安望江工业集团有限公司工程师, 南开大学计算机学院博士研究生. 主要研究方向为目标检测与跟踪和多模态感知数据融合. 本文通信作者.
E-mail: zhangyongtju@163.com
(**ZHANG Yong** Engineer at Chongqing Chang′an Wangjiang Industry Co., Ltd., and Ph.D. candidate at the College of Computer Science, Nankai University. His research interest covers object detection and tracking, and multimodal perception data fusion. Corresponding author of this paper.)

韩树豪 南开大学计算机学院硕士研究生. 主要研究方向为深度学习和视频插帧.
E-mail: hansh@mail.nankai.edu.cn
(**HAN Shu-Hao** Master student at the College of Computer Science, Nankai University. His research interest covers deep learning and video frame interpolation.)

郭春乐 南开大学计算机学院副教授, 南开国际先进研究院 (深圳福田) 副教授. 主要研究方向为计算成像, 图像增强与复原.
E-mail: guochunle@nankai.edu.cn
(**GUO Chun-Le** Associate professor at the College of Computer Science, Nankai University, and Nankai International Advanced Research Institute (SHENZHEN-FUTIAN). His research interest covers computational imaging, image enhancement and restoration.)

李重仪 南开大学计算机学院教授, 南开国际先进研究院 (深圳福田) 教授. 主要研究方向为计算成像.
E-mail: lichongyi@nankai.eud.cn
(**LI Chong-Yi** Professor at the College of Computer Science, Nankai University, and Nankai International Advanced Research Institute (SHENZHEN-FUTIAN). His main research interest is computational imaging.)

程明明 南开大学计算机学院教授, 南开国际先进研究院 (深圳福田) 教授. 主要研究方向为人工智能, 计算机视觉和计算机图形学.
E-mail: cmm@nankai.edu.cn
(**CHENG Ming-Ming** Professor at the College of Computer Science, Nankai University, and Nankai International Advanced Research Institute (SHENZHEN-FUTIAN). His research interest covers artificial intelligence, computer vision and computer graphics.)