基于深度学习的视频超分辨率重建算法进展

唐 麒[1,2] 赵 耀[1,2] 刘美琴[1,2] 姚 超[3]

摘 要 视频超分辨率重建 (Video super-resolution, VSR) 是底层计算机视觉任务中的一个重要研究方向, 旨在利用低分辨率视频的帧内和帧间信息, 重建具有更多细节和内容一致的高分辨率视频, 有助于提升下游任务性能和改善用户观感体验. 近年来, 基于深度学习的视频超分辨率重建算法如雨后春笋般涌现, 在帧间对齐、信息传播等方面取得了突破性的进展. 在简述视频超分辨率重建任务的基础上, 梳理了现有的视频超分辨率重建的公共数据集及相关算法; 接着, 重点综述了基于深度学习的视频超分辨率重建算法的创新性工作进展情况; 最后, 总结了视频超分辨率重建算法面临的挑战及未来的发展趋势.

关键词 视频超分辨率重建,深度学习,循环神经网络,注意力机制,光流估计,可变形卷积

引用格式 唐麒, 赵耀, 刘美琴, 姚超. 基于深度学习的视频超分辨率重建算法进展. 自动化学报, 2025, **XX**(X): X−X

DOI 10.16383/j.aas.c240235

A Review of Video Super-resolution Algorithms Based on Deep Learning

TANG Qi[1,2] ZHAO Yao[1,2] LIU Mei-Qin[1,2] YAO Chao[3]

Abstract Video super-resolution (VSR) is an essential research realm within low-level vision tasks. It aims to reconstruct high-resolution video with realistic details and coherent content by utilizing intra-frame and inter-frame information of low-resolution video, which positively impacts the performance of downstream tasks and the improvement of user's perception experience. In recent years, VSR base on deep learning has emerged abundantly. These methods have continuously exploited and broken through from perspective such as inter-frame alignment and information propagation. On the basis of briefly describing the task of VSR, the existing public data sets and related algorithms are combed. Subsequently, the focus shifts to the innovative work progress of deep-learning-based VSR. Finally, the challenges and future development trends of VSR algorithms are outlined.

Key words Video super-resolution, deep learning, recurrent neural network, attention, optical flow estimation, deformable convolution

Citation Tang Qi, Zhao Yao, Liu Mei-Qin, Yao Chao. A review of video super-resolution algorithms based on deep learning. *Acta Automatica Sinica*, 2025, **XX**(X): X−X

随着便携式消费级相机的发展和第五代移动通信技术的普及, 视频已成为人们日常生活中最主要的视觉媒介之一. 在用户追求高清画质的同时, 相机拍摄得到的视频受到采集设备精度、网络传输带宽等因素的制约, 造成视频的成像分辨率和采样频率低、存在噪声等复杂的退化问题. 由图像超分辨率重建延伸而来的视频超分辨率重建 (Video super-resolution, VSR) 旨在将低分辨率 (Low-resolution, LR) 视频重建为相应的高分辨率 (High-resolution, HR) 版本. 视频超分既要求复原的高分辨率视频与给定的低分辨率视频保持内容的一致性和连续性, 同时又希望重建的高分辨率视频细节清晰、真实和自然, 广泛应用于影像修复[1]、网络传输[2]和智能分析[3–4]等领域.

视频超分辨率重建算法可以利用相邻帧包含的高度相关的时序信息, 但时序信息未对齐会导致重建的视频出现帧间内容不连贯、抖动的现象, 限制了视频超分辨率重建性能的提升. 因此, 如何有效利用视频的帧间信息是视频超分辨率重建研究的热点和难点. 早期的 VSR 算法直接将图像超分的插值法 (如双线性插值、双立方插值) 应用于每帧视频的高分辨率重建, 并未考虑视频的时间维度造成重建视频出现帧间不连贯的问题. 因此, 有些算法结合贝叶斯[5]、最大期望[6]等技术和视频的时空信息提升视频的重建质量. 然而, 这些方法受限于严苛的假设条

收稿日期 2024-04-29 录用日期 2024-10-16
Manuscript received April 29, 2024; accepted October 16, 2024
中央高校基本科研业务费专项资金资助 (2024JBZX001), 国家自然科学基金 (62120106009, 62332017, 62372036) 资助
Supported by Fundamental Research Funds for the Central Universities (2024JBZX001), and National Natural Science Foundation of China (62120106009, 62332017, 62372036)
本文责任编委 左旺孟
Recommended by Associate Editor Zuo Wang-Meng
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 3. 北京科技大学计算机与通信工程学院 北京 100083
1. Institute of Information Science, Beijing Jiaotong University, Beijing 100044 2. Visual Intelligence + X International Cooperation Joint Laboratory of Ministry of Education, Beijing Jiaotong University, Beijing 100044 3. School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083

件, 不足以适应复杂多变的视频内容. 此外, 这些方法难以有效复原丢失的高频信息, 无法满足目前高清显示设备的需求.

伴随着深度学习技术在图像处理领域的成功应用[7–8], 基于深度学习的视频超分辨率重建算法也迅速发展. 得益于强大的非线性学习能力, 基于深度学习的 VSR 算法可以有效提取和融合视频中的时空信息, 获得了优于传统 VSR 算法的重建效果. 基于深度学习的 VSR 算法主要包括对齐、融合和重建三个部分. 其中, 对齐模块利用光流、可变形卷积等方法可以解决卷积操作导致的感受野受限、应对相邻帧中较大运动变化造成的模糊和伪影问题. 基于循环神经网络的方法利用了之前视频帧信息的隐状态完成视频帧的对齐和融合, 克服了卷积神经网络无法建模长程时序信息的缺陷. 基于 Transformer 的 VSR 算法利用注意力机制获取视频帧内和帧间的相关性, 获得了优于其他网络结构的特征表示能力和视频的重建性能, 可以并行处理所有视频帧, 不存在循环网络的特征衰减和噪声放大等问题. 上述算法主要将 VSR 视作回归问题, 在大规模视频序列数据集上学习从低分辨率到高分辨率视频帧的非线性映射, 但重建的视频帧中存在纹理模糊的问题. 基于生成模型的 VSR 算法可以在保持重建视频时序一致性的同时, 有效地建模高分辨率视频帧的数据分布, 这类方法复原的高分辨率视频中包含了清晰、真实的细节.

本文聚焦于基于深度学习的 VSR 算法, 从研究进展与存在的问题及挑战等方面全面梳理了 VSR , 并系统概述了相关技术的进展情况.

1 视频超分辨率重建

由于设备成像能力的限制和采集环境的噪声干扰, 真实场景下拍摄得到的视频通常存在着分辨率低、噪声以及模糊等问题. 视频超分辨率重建的观测模型可以表示为:

$$I_t^{\mathrm{LR}} = \mathcal{D}(I_t^{\mathrm{HR}}; \delta) \tag{1}$$

其中, I_t^{LR} 和 I_t^{HR} 分别表示视频序列中第 t 帧的低分辨率版本和高分辨率版本, $\mathcal{D}(\cdot)$ 描述下采样、模糊等退化映射, δ 表示各种退化模型的参数, 如下采样因子、噪声因子和模糊因子等. 视频超分辨率重建以连续的 $2N+1$ 帧作为输入, 利用视频的帧内和帧间信息, 由退化的低分辨视频帧 I_t^{LR} 重建出高分辨率视频帧 I_t^{SR}, 要求 I_t^{SR}尽可能地接近真实的视频帧 I_t^{HR}, 即求解退化过程的逆过程, 可以表示为:

$$I_t^{\mathrm{SR}} = \mathcal{D}^{-1}\left(I_t^{\mathrm{LR}}, \{I_n^{\mathrm{LR}}\}_{n=t-N}^{t+N}; \theta\right) \tag{2}$$

其中, I_t^{SR} 表示重建的高分辨视频的第 t 帧, θ 表示超分重建模型的参数. 显然超分辨率重建是一个典型的不适定 (ill-posed) 问题, 即对于给定的低分辨率视频重建的高分辨率视频并不唯一. 为了使重建的视频与真实的高分辨率视频接近, 基于深度学习的 VSR 选择不同的损失函数优化参数 $\widehat{\theta}$, 如像素损失、内容损失和对抗损失等, 即基于深度学习视频超分辨率重建目标可以表示为:

$$\widehat{\theta} = \arg\min_{\theta} \mathcal{L}\left(I^{\mathrm{SR}}, I^{\mathrm{HR}}\right) \tag{3}$$

其中, $\widehat{\theta}$ 表示优化得到的 VSR 模型参数, $\mathcal{L}(\cdot,\cdot)$ 用于度量重建的高分辨率视频帧和真实高分辨率视频帧之间的差异. 常用的损失函数包括 L1 损失、L2 损失和 Charbonnier 损失[9] $\mathcal{L}_{\mathrm{Char}}$. 其中, $\mathcal{L}_{\mathrm{Char}}$ 有助于 VSR 在保留重要视觉细节的同时也抑制了噪声的影响、提高了 VSR 的性能[10–11], 可以表示为:

$$\mathcal{L}_{\mathrm{Char}} = \frac{1}{N}\sum_{i=0}^{N-1}\sqrt{\left(I_i^{\mathrm{SR}} - I_i^{\mathrm{HR}}\right)^2 + \epsilon^2} \tag{4}$$

其中, N 表示视频帧的像素数量, ϵ 表示一个非常小的常量. 在参数空间寻找最优参数离不开大量低-高分辨率视频对的支持, 图 1 给出了 VSR 常用的数据集. 按照低分辨率视频的制作方式, 现有的视频超分辨率重建数据集可以分为合成数据集和真实数据集两类. 其中, 合成数据集中的低分辨率视频 I_t^{LR} 一般通过对高分辨率视频 I_t^{HR} 进行双三次插值下采样或者高斯模糊下采样得到, 定义如下:

$$I_t^{\mathrm{LR}} = I_t^{\mathrm{HR}} \downarrow_s^{\mathrm{bic}} \tag{5}$$

$$I_t^{\mathrm{LR}} = \left(I_t^{\mathrm{HR}} \otimes k\right) \downarrow_s \tag{6}$$

其中, s 和 k 分别表示下采样因子和高斯核, $\downarrow^{\mathrm{bic}}$ 和 $\otimes$ 分别表示下采样和卷积操作. 常用的 VSR 数据集信息如表 1 所示, Vimeo-90K[12]视频序列规模大常用作训练集, Vid4[13]因视频序列高频细节丰富常作为测试集, REDS[14]则因视频序列较长且对象的运动幅度大而被视为更具挑战的视频超分数据集.

除此之外, 越来越多的研究开始关注真实场景的视频超分辨率重建. 与合成数据集相比, 真实场景中采集的视频退化过程复杂且未知. 为解决现有基于深度学习的视频超分方法在真实场景中性能显著退化的问题, 研究人员在设计更加鲁棒的算法的同时, 也致力于制作在真实场景中采集低-高分辨率的视频数据集. 按照训练数据制作方式的不同, 真实场景视频超分数据集可以分为利用不同焦距的相机采集的数据集和利用高阶退化模型合成的数据集.

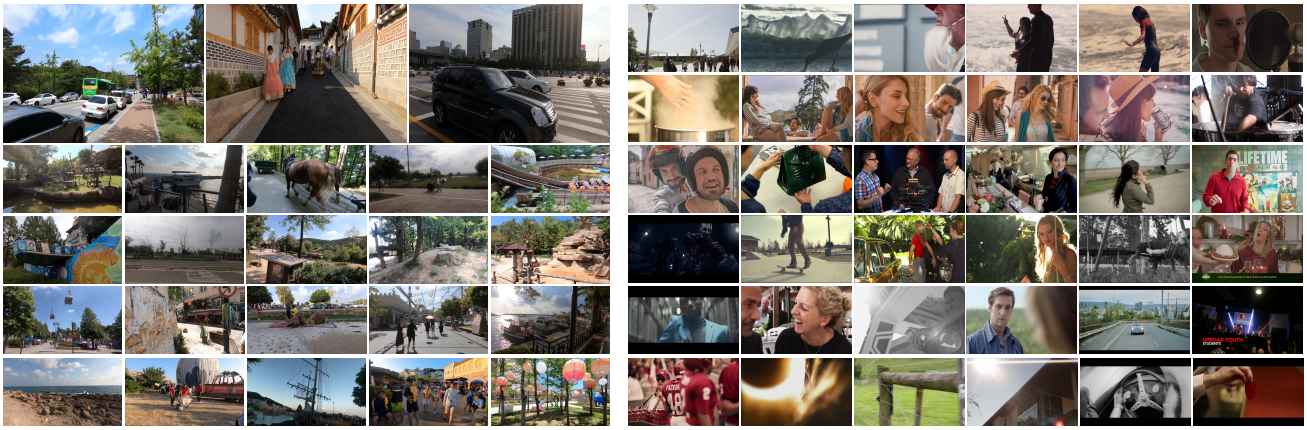

图 1　视频超分辨率重建数据集 REDS (左) 和 Vimeo-90K (右) 示例

Fig. 1　Examples of video super-resolution datasets from REDS (left) and Vimeo-90K (right)

表1　基于深度学习的视频超分辨率重建数据集

Table 1　Datasets of video super-resolution based on deep learning

	数据集		类型	视频数量	帧数	分辨率	颜色空间
合成数据集	YUV25 [15]		训练集	25	-	386 × 288	YUV
	TDTFF [16]	Turbine	测试集	5	-	648 × 528	YUV
		Dancing				950 × 530	
		Treadmill				700×600	
		Flag				1000 × 580	
		Fan				990 × 740	
	Vid4 [13]	Foliage	测试集	4	49	720 × 480	RGB
		Walk			47	720 × 480	
		Calendar			41	720 × 576	
		City			34	704 × 576	
	YUV21 [17]		测试集	21	100	352 × 288	YUV
	Venice [18]		训练集	1	1 077	3 840×2 160	RGB
	Myanmar [19]		训练集	1	527	3 840×2 160	RGB
	CDVL [20]		训练集	100	30	1 920×1 080	RGB
	UVGD [21]		测试集	16	-	3 840×2 160	YUV
	LMT [22]		训练集	26	-	1 920×1 080	YCbCr
	SPMCS [23]		训练集和测试集	975	31	960×540	RGB
	MM542 [24]		训练集	542	32	1 280×720	RGB
	UDM10 [25]		测试集	10	32	1 272×720	RGB
	Vimeo-90K [12]		训练集和测试集	91 701	7	448×256	RGB
	REDS [14]		训练集和测试集	270	100	1 280×720	RGB
	Parkour [26]		测试集	14	-	960×540	RGB
真实数据集	RealVSR [27]		训练集和测试集	500	50	1 024×512	RGB/YCbCr
	VideoLQ [28]		测试集	50	100	1 024×512	RGB
	RealMCVSR [29]		训练集和测试集	161	-	1 920×1 080	RGB
	MVSR4× [30]		训练集和测试集	300	100	1 920×1 080	RGB
	DTVIT [31]		训练集和测试集	196	100	1 920×1 080	RGB
	YouHQ [32]		训练集和测试集	38 616	32	1 920×1 080	RGB

例如, 为了模拟真实场景的复杂退化过程, Real-BasicVSR [28] 在训练阶段采用了二阶退化模型来生成复杂且退化过程未知的低质量视频. Real-BasicVSR 采用的退化操作可以分为基于图像的退化操作和基于视频的退化操作. 具体地, 在图像层面, Real-BasicVSR 遵循 Real-ESRGAN [33] 的设计, 将随机模糊 (如高斯滤波、sinc 滤波等)、尺寸调整 (如双立方插值、双线性插值等)、噪声 (如高斯噪声、泊

松噪声等) 和 JPEG 压缩作为基于图像的退化操作. 在视频层面, Real-BasicVSR 额外引入了视频压缩技术, 考虑了视频帧间的隐式依赖关系. 对于视频压缩, Real-BasicVSR 随机选取“libx264”、“h264”和“mpeg4”中的一种编解码器和 $[10^4, 10^5]$ 范围内比特率. 因此, Real-BasicVSR 的退化模型综合引入图像层面和视频层面的退化方式, 在时间和空间上提供了复杂且随机的退化过程.

与图像超分辨率重建一样, 视频超分辨率重建采用峰值信噪比 (Peak Signal-to-Noise Ratio, PSNR) 来度量重建帧和高分辨率帧之间的像素差异, PSNR 依赖于均值误差 (Mean Squared Error, MSE), 可以表示为:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(I_i^{\mathrm{HR}} - I_i^{\mathrm{SR}}\right)^2 \tag{7}$$

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{{I_{\max}^{\mathrm{HR}}}^2}{\mathrm{MSE}}\right) \tag{8}$$

其中, I_i^{HR} 和 I_i^{SR} 分别表示真实的高分辨视频帧和重建的高分辨率视频帧. $I_{\max}^{\mathrm{HR}}$ 表示真实的高分辨率视频帧中最大的像素值. PSNR 值越高意味着重建图像质量越好, 即失真越小. 然而, PSNR 在捕捉感知差异方面存在着一定的局限性, 因此 PSNR 值并不与人类的主观感知质量密切相关. 结构相似性 (Structural similarity index measure, SSIM) 衡量两幅图像亮度、对比度、结构三方面之间的相似性, 可以表示为:

$$\mathrm{SSIM} = \frac{(2\mu_{I^{\mathrm{HR}}}\mu_{I^{\mathrm{SR}}} + c_1)(2\sigma_{I^{\mathrm{HR}},I^{\mathrm{SR}}} + c_2)}{(\mu_{I^{\mathrm{HR}}}^2 + \mu_{I^{\mathrm{SR}}}^2 + c1)(\sigma_{I^{\mathrm{HR}}}^2 + \sigma_{I^{\mathrm{SR}}}^2 + c_2)} \tag{9}$$

其中, $\mu_{I^{\mathrm{HR}}}$ 和 $\mu_{I^{\mathrm{SR}}}$ 分别表示真实和重建视频帧的均值, $\sigma_{I^{\mathrm{HR}}}^2$ 和 $\sigma_{I^{\mathrm{SR}}}^2$ 分别表示真实和重建视频帧的方差, $\sigma_{I^{\mathrm{HR}},I^{\mathrm{SR}}}$ 表示真实和重建视频帧的协方差, c_1 和 c_2 为常数. SSIM 的值介于 0 和 1 之间, 其值越高, 重建视频的质量越好. 此外, 学习的感知图像块相似性 (Learned perceptual image patch similarity, LPIPS)[34]和时间一致性 (Temporal motion-based video integrity evaluation index, T-Movie)[6]也可以用于评价视频的重建质量.

不同于基于像素差异的指标, LPIPS 用于测量两幅图像之间感知相似性, 利用预训练的深度神经网络所提取的图像特征差异来衡量图像之间的差异, 可以模拟人类的视觉感知. LPIPS 值越低表示两张图像越相似, 可以表示为:

$$\mathrm{LPIPS} = \sum_{l,h,w}\frac{\left\|w_l \odot \left(\mathrm{VGG}_\theta^l(I_{hw}^{\mathrm{SR}}) - \mathrm{VGG}_\theta^l(I_{hw}^{\mathrm{HR}})\right)\right\|_2}{H_l W_l} \tag{10}$$

其中, $\mathrm{VGG}_\theta^l(\cdot)$ 表示利用在大规模数据集训练的 VGG 网络获得的第 l 层特征, 能够捕捉视频帧的高级语义信息. 然而, 基于分数的质量评价无法刻画图像复杂的局部性和内容相关性, 限制了更深层次的质量感知. 随着大语言模型和多模态语言模型的出现, 语言成为描述视频感知质量的工具. 例如, DepictQA[35] 是一种基于多模态语言模型的图像质量感知方法, 可以对视频帧质量进行类似于人类的、基于语言的描述. 视频超分辨率重建任务也亟待类似的评价方法, 更加全面地评价重建视频的感知质量和时序一致性[36–39].

目前, 视频超分辨率重建领域仍采用 PSNR 和 SSIM 作为主要评价指标. 当上采样因子为 4× 时, 低分辨率视频通过双三次插值完成 4 倍下采样操作, 代表性的 VSR 算法重建视频的 PSNR 和 SSIM 如表 2 所示. 当上采样因子为 4× 时, 采用标准差为 $\sigma = 1.6$ 的滤波器进行高斯模糊后再进行 4 倍下采样操作, 典型的 VSR 的 PSNR 和 SSIM 如表 3 所示. VSR 算法在 REDS4 数据集上评估 RGB 通道的性能, 在其他数据集评估 Y 通道的性能. 通常, 模型在重建过程中可以利用的时序信息越多, 评估其生成高分辨率视频的评价指标就越高 (如 PSRT-sliding 和 PSRT-recurrent). 在参考相同时序信息的情况下, 更强的特征提取模块和更精确的帧间对齐模块可以进一步提升 VSR 模型的视频重建性能 (如 MFPI 和 IART). 代表性的 VSR 算法重建视频的可视化结果如图 2 和图 3 所示.

相较于基于合成数据集的视频超分方法, 真实场景的视频超分算法仍处于早期探索阶段. 虽然相关工作为了解决真实场景的视频超分问题已经制作了一系列的数据集, 但由于缺乏统一的标准, 现有的真实场景视频超分算法面临着训练数据集和评价标准不一致的问题. 这一现状对公平对比现有算法以及推动未来研究的发展提出了严峻的挑战. 基于现有的研究工作, 在 RealVSR 和 MVSR 4× 两个真实场景数据集的性能对比结果如表 4 所示. 尽管基于合成数据集的视频超分算法已经形成了较为固定的测评标准, 但由于规模小、分辨率低等问题, 导致这些数据集难以满足不断增长的模型参数量对大规模视频数据的需求. 随着扩散模型在视频生成领域的不断发展及其在真实场景单图超分任务的出色表现, 大规模、高质量的视频数据也成为推动基于扩散模型的视频超分发展的充分条件.

表2 对双三次插值下采样后的视频进行 VSR 的性能对比结果

Table 2 Performance comparison of video super-resolution algorithm with bicubic downsampling

对比方法	训练帧数	参数量 (M)	双三次插值下采样		
			REDS (RGB 通道)	Vimeo-90K-T (Y 通道)	Vid4 (Y 通道)
Bicubic	-	-	26.14/0.7292	31.32/0.8684	23.78/0.6347
VSRNet [40]	-	0.27	-/-	-/-	22.81/0.6500
VSRResFeatGAN [41]	-	-	-/-	-/-	24.50/0.7023
VESPCN [42]	-	-	-/-	-/-	25.35/0.7577
VSRResNet [41]	-	-	-/-	-/-	25.51/0.7530
SPMC [23]	-	2.17	-/-	-/-	25.52/0.7600
3DSRNet [43]	-	-	-/-	-/-	25.71/0.7588
RRCN [44]	-	-	-/-	-/-	25.86/0.7591
TOFlow [12]	5/7	1.41	27.98/0.7990	33.08/0.9054	25.89/0.7651
STARNet [45]	-	111.61	-/-	30.83/0.9290	-/-
MEMC-Net [46]	-	-	-/-	33.47/0.9470	24.37/0.8380
STMN [47]	-	-	-/-	-/-	25.90/0.7878
SOFVSR [48]	-	1.71	-/-	-/-	26.01/0.7710
RISTN [49]	-	3.67	-/-	-/-	26.13/0.7920
MMCNN [24]	-	10.58	-/-	-/-	26.28/0.7844
RTVSR [50]	-	15.00	-/-	-/-	26.36/0.7900
TDAN [51]	-	1.97	-/-	-/-	26.42/0.7890
D3DNet [52]	-/7	2.58	-/-	35.65/0.9330	26.52/0.7990
FFCVSR [53]	-	-	-/-	-/-	26.97/0.8300
EVSRNet [54]	-	-	27.85/0.8000	-/-	-/-
StableVSR [55]	-	-	27.97/0.8000	-/-	-/-
DUF [56]	7/7	5.8	28.63/0.8251	-/-	27.33/0.8319
PFNL [57]	7/7	3	29.63/0.8502	36.14/0.9363	26.73/0.8029
DNSTNet [58]	-	-	-/-	36.86/0.9387	27.21/0.8220
RBPN [59]	7/7	12.2	30.09/0.8590	37.07/0.9435	27.12/0.8180
DSMC [60]	-	11.58	30.29/0.8381	-/-	27.29/0.8403
Boosted EDVR [31]	-	-	30.53/0.8699	-/-	-/-
TMP [61]	-	3.1	30.67/0.8710	-/-	27.10/0.8167
MuCAN [62]	5/7	-	30.88/0.8750	37.32/0.9465	-/-
MSFFN [63]	-	-	-/-	37.33/0.9467	27.23/0.8218
DAP [64]	15/5	-	30.59/0.8703	-/-	-/-
MultiBoot VSR [65]	-	60.86	31.00/0.8822	-/-	-/-
SSL-bi [66]	15/14	1.0	31.06/0.8933	36.82/0.9419	27.15/0.8208
EDVR [67]	5/7	20.6	31.09/0.8800	37.61/0.9489	27.35/0.8264
RLSP [68]	-	4.2	-/-	37.39/0.9470	27.15/0.8202
TGA [69]	-	5.8	-/-	37.43/0.9480	27.19/0.8213
KSNet-bi [70]	-	3.0	31.14/0.8862	37.54/0.9503	27.22/0.8245
VSR-T [71]	5/7	32.6	31.19/0.8815	37.71/0.9494	27.36/0.8258
PSRT-sliding [72]	5/-	14.8	31.32/0.8834	-/-	-/-
SeeClear [73]	5/5	229.23	31.32/0.8856	37.64/0.9503	27.80/0.8404
DPR [74]	-	6.3	31.38/0.8907	37.11/0.9446	27.19/0.8243
BasicVSR [75]	15/14	6.3	31.42/0.8909	37.18/0.9450	27.24/0.8251
Boosted BasicVSR [31]	-	-	31.42/0.8917	-/-	-/-
SATeCo [76]	6/6	-	31.62/0.8932	-/-	27.44/0.8420
IconVSR [75]	15/14	8.7	31.67/0.8948	37.47/0.9476	27.39/0.8279
ICNet [77]	-	18.34	31.71/0.8963	37.72/0.9477	27.43/0.8287
MSHPFNL [78]	-	7.77	-/-	36.75/0.9406	27.70/0.8472

表2 对双三次插值下采样后的视频进行 VSR 的性能对比结果 (续)

Table 2 Performance comparison of video super-resolution algorithm with bicubic downsampling (Continued)

对比方法	训练帧数	参数量 (M)	双三次插值下采样		
			REDS (RGB 通道)	Vimeo-90K-T (Y 通道)	Vid4 (Y 通道)
PA [79]	5/7	38.2	32.05/0.8941	-/-	28.02/0.8373
FTVSR [80]	-	10.8	31.82/0.8960	-/-	-/-
C^2-Matching [81]	-	-	32.05/0.9010	-/-	**28.87**/**0.8960**
ETDM [82]	-	8.4	32.15/0.9024	-/-	-/-
BasicVSR++ [83]	30/14	7.3	32.39/0.9069	37.79/0.9500	27.79/0.8400
RTA [84]	5/7	17	31.30/0.8850	37.84/0.9498	27.90/0.8380
Semantic Lens [85]	5/-	-	31.42/0.8881	-/-	-/-
TCNet [86]	-	9.6	31.82/0.9002	37.94/0.9514	27.48/0.8380
TTVSR [87]	50/-	6.8	32.12/0.9021	-/-	-/-
VRT [88]	16/7	35.6	32.19/0.9006	38.20/0.9530	27.93/0.8425
CTVSR [89]	16/14	34.5	32.28/0.9047	-/-	28.03/0.8487
FTVSR++ [90]	-	10.8	32.42/0.9070	-/-	-/-
LGDFNet-BPP [91]	-	9.0	32.53/0.9007	-/-	27.99/0.8409
PP-MSVSR-L [92]	-	7.4	32.53/0.9083	-/-	-/-
CFD-BasicVSR++ [127]	30/7	7.5	32.51/0.9083	37.90/0.9504	27.84/0.8406
RVRT [93]	30/14	10.8	32.75/0.9113	38.15/0.9527	27.99/0.8426
DFVSR [94]	-	7.1	32.76/0.9081	38.25/**0.9556**	27.92/0.8427
PSRT-recurrent [72]	16/14	13.4	32.72/0.9106	38.27/0.9536	28.07/0.8485
MFPI [95]	-/-	7.3	32.81/0.9106	**38.28**/0.9534	28.11/0.8481
EvTexture [96]	15/-	8.9	32.79/0.9174	38.23/0.9544	29.51/0.8909
MIA-VSR [97]	16/14	16.5	32.78/**0.9220**	38.22/0.9532	28.20/0.8507
CFD-PSRT [127]	30/7	13.6	32.83/0.9140	38.33/0.9548	28.18/0.8503
IART [98]	16/7	13.4	**32.90**/0.9138	38.14/0.9528	28.26/0.8517
EvTexture+ [96]	15/-	10.1	**32.93**/**0.9195**	**38.32**/ **0.9558**	**29.78**/**0.8983**

表3 对高斯模糊下采样后的视频进行 VSR 的性能对比结果

Table 3 Performance comparison of video super-resolution algorithm with blur downsampling

对比方法	训练帧数	参数量 (M)	高斯模糊下采样		
			UDM10 (Y 通道)	Vimeo-90K-T (Y 通道)	Vid4 (Y 通道)
Bicubic	-	-	28.47/0.8253	31.30/0.8687	21.80/0.5246
BRCN [99]	-	-	-/-	-/-	24.43/0.6334
ToFNet [12]	5/7	1.41	36.26/0.9438	34.62/0.9212	25.85/0.7659
TecoGAN [100]	-	3.00	-/-	-/-	25.89/-
SOFVSR [48]	-	1.71	-/-	-/-	26.19/0.7850
RRN [101]	-	3.4	38.96/0.9644	-/-	27.69/0.8488
TDAN [51]	-	1.97	-/-	-/-	26.86/0.8140
FRVSR [102]	-	5.1	-/-	-/-	26.69/0.8220
DUF [56]	7/7	5.8	38.48/0.9605	36.87/0.9447	27.38/0.8329
RLSP [68]	-	4.2	38.48/0.9606	36.49/0.9403	27.48/0.8388
PFNL [57]	7/7	3	38.74/0.9627	-/-	27.16/0.8355
RBPN [59]	7/7	12.2	38.66/0.9596	37.20/0.9458	27.17/0.8205
TMP [61]	-	3.1	-/-	37.33/0.9481	27.61/0.8428
TGA [69]	-	5.8	38.74/0.9627	37.59/0.9516	27.63/0.8423
SSL-bi [66]	15/14	1.0	39.35/0.9665	37.06/0.9458	27.56/0.8431
RSDN [103]	-	6.19	-/-	37.23/0.9471	27.02/0.8505
DAP [64]	15/5	-	39.50/0.9664	37.25/0.9472	-/-

表3　对高斯模糊下采样后的视频进行 VSR 的性能对比结果 (续)

Table 3　Performance comparison of video super-resolution algorithm with blur downsampling (Continued)

对比方法	训练帧数	参数量 (M)	高斯模糊下采样		
			UDM10 (Y 通道)	Vimeo-90K-T (Y 通道)	Vid4 (Y 通道)
SeeClear [73]	5/5	229.23	39.72/0.9675	-/-	-/-
EDVR [67]	5/7	20.6	39.89/0.9686	37.81/0.9523	27.85/0.8503
DPR [74]	-	6.3	39.72/0.9684	37.24/0.9461	27.89/0.8539
BasicVSR [75]	15/14	6.3	39.96/0.9694	37.53/0.9498	27.96/0.8553
IconVSR [75]	15/14	8.7	40.03/0.9694	37.84/0.9524	28.04/0.8570
R2D2 [104]	-	8.25	39.53/0.9670	-/-	28.13/**0.9244**
FTVSR [80]	-	10.8	-/-	-/-	28.31/0.8600
FDAN [105]	-	-	39.91/0.9686	37.75/0.9522	27.88/0.8508
PP-MSVSR [92]	-	1.45	40.06/0.9699	37.54/0.9499	28.13/0.8604
GOVSR [106]	-	-	40.14/0.9713	37.63/0.9503	28.41/0.8724
ETDM [82]	-	8.4	40.11/0.9707	-/-	28.81/0.8725
TTVSR [87]	50/-	6.8	40.41/0.9712	37.92/0.9526	28.40/0.8643
BasicVSR++ [83]	30/14	7.3	40.72/0.9722	38.21/0.9550	29.04/0.8753
CFD-BasicVSR++ [127]	30/7	7.5	40.77/0.9726	38.36/0.9557	29.14/0.8760
TCNet [86]	-	9.6	-/-	-/-	28.44/0.8730
VRT [88]	16/7	35.6	41.05/0.9737	**38.72**/**0.9584**	29.42/0.8795
CTVSR [89]	16/14	34.5	**41.20**/0.9740	**38.83**/**0.9580**	29.28/0.8811
FTVSR++ [90]	-	10.8	-/-	-/-	28.80/0.8680
LGDFNet-BPP [91]	-	9.0	40.81/**0.9756**	-/-	29.39/0.8798
RVRT [93]	30/14	10.8	40.90/0.9729	38.59/0.9576	**29.54**/0.8810
DFVSR [94]	-	7.1	40.97/0.9733	38.51/0.9571	**29.56**/**0.8983**
MFPI [95]	-/-	7.3	**41.08**/**0.9741**	38.70/0.9579	29.34/0.8781

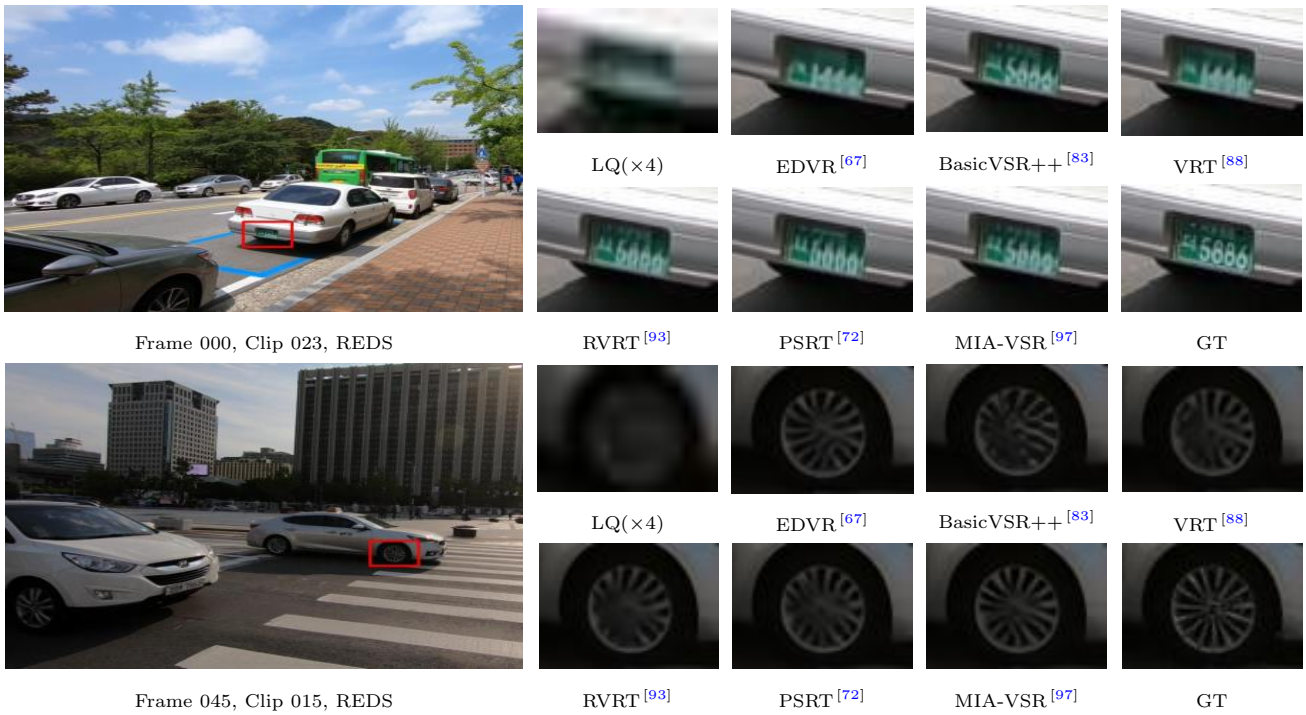

图 2　部分 VSR 模型在 REDS 据集的可视化比较结果

Fig. 2　Visual comparison results of VSR methods on REDS dataset

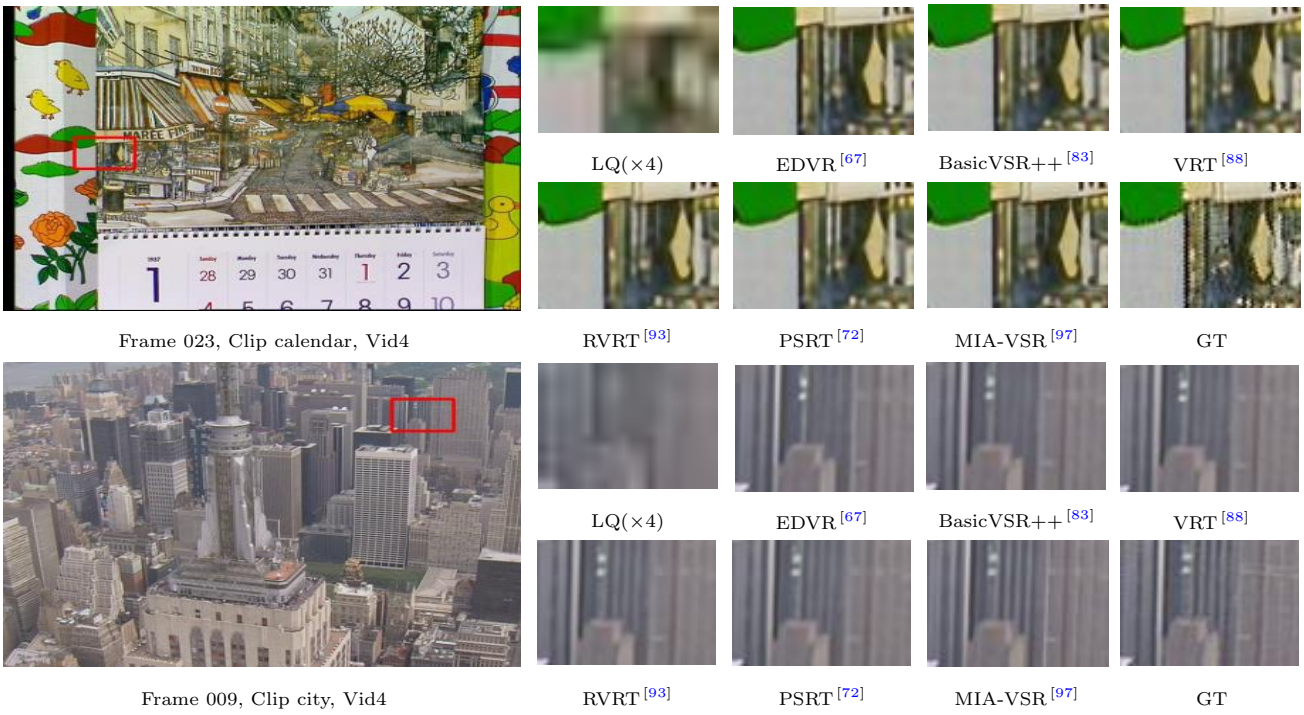

图 3 部分 VSR 模型在 Vid4 数据集的可视化比较结果

Fig. 3 Visual comparison results of VSR methods on Vid4 dataset

表4 真实场景下的 VSR 性能对比结果

Table 4 Performance comparison of real-world video super-resolution algorithm

对比方法	推理帧数	RealVSR	MVSR 4×
		PSNR/SSIM/LPIPS	PSNR/SSIM/LPIPS
RSDN [103]	之前帧	23.91/0.7743/0.224	23.15/0.7533/0.279
FSTRN [107]	7	23.36/0.7683/0.240	22.66/0.7433/0.315
TOF [12]	7	23.62/0.7739/0.220	22.80/0.7502/0.279
TDAN [51]	7	23.71/0.7737/0.229	23.07/0.7492/0.282
EDVR [67]	7	23.96/0.7781/0.216	23.51/0.7611/0.268
BasicVSR [75]	所有帧	24.00/0.7801/0.209	23.38/0.7594/0.270
MANA [108]	所有帧	23.89/0.7781/0.224	23.15/0.7513/0.285
TTVSR [87]	所有帧	24.08/0.7837/0.213	23.60/0.7686/0.277
ETDM [82]	所有帧	24.13/**0.7896**/0.206	23.61/0.7662/0.260
BasicVSR++ [83]	所有帧	24.24/0.7933/0.216	**23.70**/**0.7713**/0.263
RealBasicVSR [28]	所有帧	23.74/0.7676/**0.174**	23.15/0.7603/**0.202**
EAVSR [30]	所有帧	**24.20**/0.7862/0.208	23.61/0.7618/0.264
EAVSR+ [30]	所有帧	**24.41**/**0.7953**/0.212	**23.94**/**0.7726**/0.259
EAVSRGAN+ [30]	所有帧	23.99/0.7726/**0.170**	23.35/0.7611/**0.199**

为了有效利用视频的时序信息, VSR 网络通常会引入各种复杂的结构, 这显然给现有 VSR 的实施和扩展提出了严峻的考验. 为此, Chan 等[75]将基于深度学习的 VSR 网络划分为信息传播、帧间对齐、特征聚合和帧上采样四个部分, 在分析和改进现有模块的基础上, 构建了简单而有效的视频超分辨率重建网络 BasicVSR. 后续的研究工作大多延续这一设计, 改进 VSR 网络中的信息传播和帧间对齐模式, 以期获得更好的重建性能. 有一些文献[111–114]总结和分析了 VSR 算法, 却未能归纳总结近期基于深度学习的 VSR 算法. 此外, 考虑到帧间对齐结果对 VSR 性能的显著影响, 一些工作根据帧间信息利用方式的不同将 VSR 算法分类. 从最新的研究工作可以看出, 这种分类方法未能完备准确地划分现有的或未来的 VSR 算法, 故本文根据网络架构, 重新梳理和分析基于深度学习的 VSR 算法.

本文的整体结构如图 4 所示. 具体而言, 基于深度学习的 VSR 算法利用时序信息涉及两个方面的设计：信息传播和帧间对齐. 本文以视频重建过程中时序信息的传播方式, 将现有的基于深度学习的 VSR 算法划分为基于并行架构的算法和基于循环架构的算法. 其中, 基于并行架构的算法又可以进一步划分为基于滑动窗口的算法和基于 Transformer 的算法, 而基于循环架构的算法可以划分为基于单向传播的算法和基于双向传播的算法. 图 5 总结了本文介绍的基于深度学习的 VSR 算法的发展时间线. 不同于帧间对齐机制为标准分类的方法, 本文可以较为全面地将现有的视频超分算法进行分类, 并向后兼容, 以期为后续的研究提供新思路.

图 4　本文的结构图

Fig. 4　Architecture of the paper

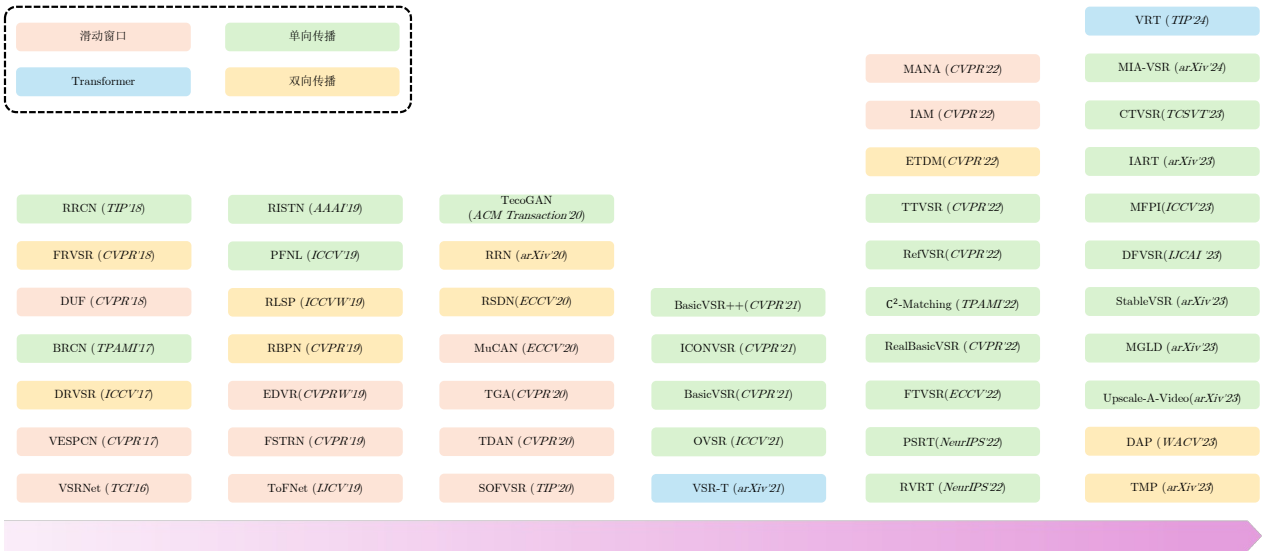

图 5　基于深度学习的视频超分辨率重建时间脉络图

Fig. 5　Timeline of video super-resolution based on deep learning

2 基于深度学习的视频超分辨率重建算法

VSR 旨在利用低分辨率视频序列蕴含的帧内空间相关性和帧间时序一致性信息重建出对应的高分辨率视频. 基于深度学习的 VSR 算法以数据驱动的方式学习从低分辨率到高分辨率视频帧的非线性映射关系, 显著地提升了重建视频帧的质量. 这些算法能够建模各种低分辨率视频的帧间时序相关性和帧内空间相关性, 以确保时间上的连贯性和空间上的一致性. 按照时序信息在传播过程中是否存在沿时间维度不断传播的隐状态, 基于深度学习的 VSR 算法可以分为基于并行架构的算法和基于循环架构的算法. 下文将系统深入地梳理、归纳和总结这两类 VSR 算法的进展情况.

2.1 基于并行架构的 VSR 算法

基于并行架构的 VSR 算法将低分辨率视频序列划分为不同的时间窗口, 不依赖于其他帧的重建信息同时提取所有帧的特征和优化视频. 该类算法以连续 N 个低分辨率视频帧作为输入, 以滑动窗口而非并行方式将局部时间窗口内的若干个支持帧用于中间目标帧的重建. 这类算法包括特征提取、特征对齐、特征融合和帧重建四个阶段. 在特征对齐阶段, 其他所有帧都与中间一帧对齐. 这种对齐方式导致 VSR 模型的计算复杂度为视频长度的平方, 复杂度高, 难以适用于较长视频序列的应用场景. 而基于 Transformer 的 VSR 算法可以同时提取、对齐和融合所有帧的特征, 达到单次重建所有帧的效果, 显著地提升了 VSR 的性能. 然而, 随之带来的巨大模型参数量和内存开销大等问题成为这类算法无法避免的缺陷.

Dong 等[115]首次将卷积神经网络引入图像超分辨率重建任务, 提出了 SRCNN 网络, 拉开了研究基于深度学习的超分辨率重建算法的序幕. 该网络训练经由低-高分辨率图像对获得了比传统 SR 算法更好的重建性能. Kappeler 等[40]将SRCNN 拓展到视频领域, 设计了视频超分辨率重建卷积神经网络 (Video super-resolution convolutional neural network, VSRNet), 结构如图 6 所示. VSRNet 包含一个三层架构, 使用在图像数据集上的预训练过程中得到的权重初始化网络参数. 在当前帧的重建过程中, VSRNet 在通道维度将过去与未来的相邻帧一同与当前帧级联以利用额外的时序信息, 并提供了三种不同的级联结构: 在网络的第一层之前级联、在网络的第一和第二层之间级联以及在网络的第二和第三层之间级联. 同时, 为解决视频中运动模糊的问题, VSRNet 采用 Druleas 算法[116]进行运动估计. 该算法基于 CLG 变分法 (Combined local-global approach with total variation) 可以获得帧间位移较大像素点的准确光流. 自适应运动补偿 (Adaptive motion compensation) 机制用于减小相邻帧的对齐误差, 减少在重建帧中的伪影. 此外, VSRNet 的滤波器对称增强模块 (Filter symmetry enforcement) 假设前后帧运动补偿误差相同, 在反向传播过程中共享对称卷积核的梯度, 从而加快了网络的训练过程.

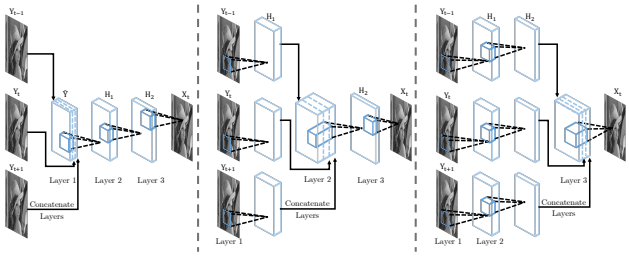

图 6 VSRNet 结构图

Fig. 6 Architecture of VSRNet

虽然 VSRNet 通过联合处理多帧的方式获得了良好的重建性能, 但其低效的运动补偿机制并不能满足实时视频超分的需求. Caballero 等[42]设计了第一个端到端的高效视频亚像素卷积网络 (Video efficient sub-pixel convolution network, VESPCN), 包括运动估计、运动补偿和时空亚像素卷积网络三个部分, 结构如图 7 所示. 其中, 运动估计模块在不同尺度的特征上由粗到细地估计光流信息, 运动补偿模块利用空间变换网络 (Spatial transformer network) 来生成两帧之间的空间变换参数后与相邻帧对齐. 在对比了不同时空相关性的处理方式后, VESPCN 采用慢融合策略 (在网络的层级结构中逐层对时序信息进行融合) 和 3D 卷积操作代替早期融合策略 (所有帧在网络第一层级联), 并使用高效无参的亚像素卷积操作进行上采样达到实时重建高质量视频帧的效果.

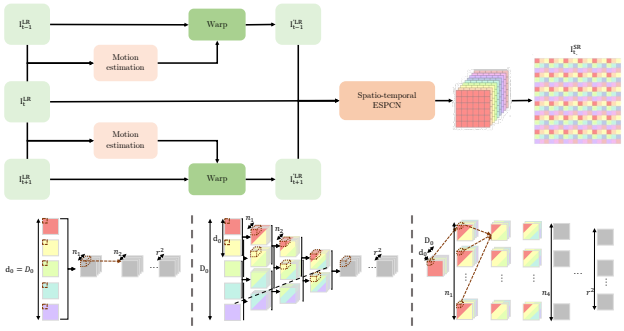

图 7 VESPCN 结构图

Fig. 7 Architecture of VESPCN

相较于低分辨率的运动信息, 高分辨率的光流信息可以建立更精细的帧间对应关系. 如图 8 所

示, Wang 等[48]提出了基于光流超分的视频超分网络 (Super-resolve optical flow for video super-resolution, SOFVSR), 利用光流重建网络 (Optical flow reconstruction network, OFRNet) 以由粗到细的方式从输入的低分辨率视频中获取高分辨率的光流信息. 具体地, OFRNet 采用卷积层和稠密残差块提取两倍下采样后的低分辨率光流, 然后经由亚像素卷积层将其上采样到高分辨率空间, 通过空间-深度间的维度变换, 高分辨率的光流又被反投影至低分辨率空间, 用于对齐低分辨率的视频帧, 再将待重建帧和对齐帧输入重建网络获得高分辨率的重建视频.

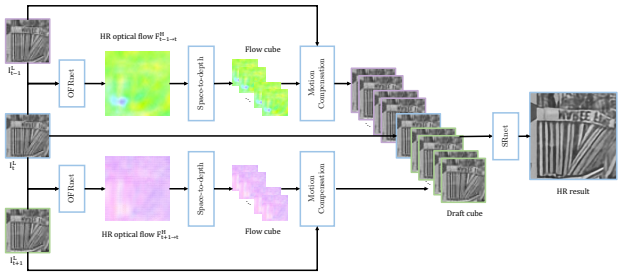

图 8　SOFVSR 结构图

Fig. 8　Architecture of SOFVSR

Xue 等[12]在端到端的训练网络过程中, 设计光流估计模块学习与视频复原任务特征表示最相关的光流, 并构建了面向任务的光流网络 (Task-oriented flownet, TOFlow), 结构如图 9 所示. 具体地, TOFlow 首先利用 SpyNet 由粗到细地估计相邻视频帧间的光流信息, 以应对帧间较大的像素位移. 然后, 空间变换网络将相邻帧对齐到当前待重建帧, 再由图像处理模块重建高分辨率的视频帧.

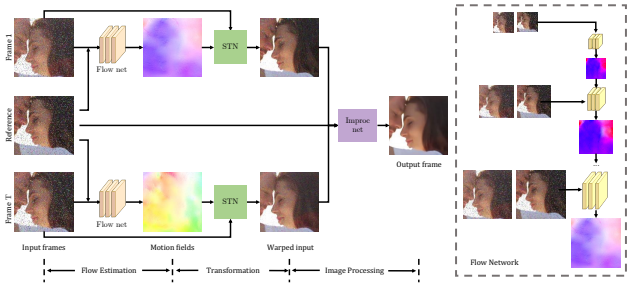

图 9　TOFlow 结构图

Fig. 9　Architecture of TOFlow

基于光流的显式运动补偿效果高度依赖于运动估计的准确程度. 当存在遮挡或复杂运动情况时, 运动信息估计不准确会导致对齐结果失真, 降低了重建帧的质量. 诸如光流之类的逐像素运动估计常常伴随着沉重的计算负荷, 显著增加了重建视频的计算开销. 受动态卷积网络的启发, Jo 等[56]提出了依赖于输入帧的动态上采样网络 (Dynamic upsampling filters, DUF), 以实现自适应的上采样操作, 结构如图 10 所示. 具体地, DUF 在完成上采样的同时也提取特征, 并结合 3D 卷积完成时空信息的建模, 避免了显式的运动估计和运动补偿. DUF 还估计了待重建帧的残差图, 进一步增强了重建结果中的高频细节.

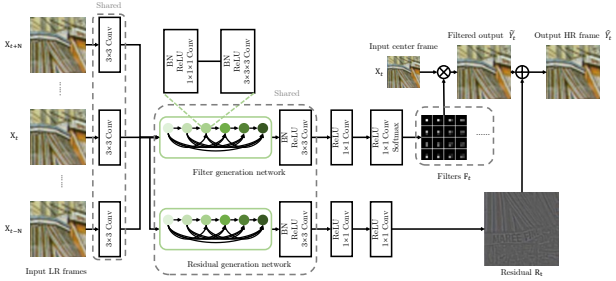

图 10　DUF 结构

Fig. 10　Architecture of DUF

与 2D 卷积相比, 3D 卷积无疑会带来更大的计算开销. Li 等[107]提出了一种快速空时残差网络 (Fast spatio-temporal residual network, FSTRN), 在建模时空依赖关系的同时提升网络的计算效率, 结构如图 11 所示. 具体地, FSTRN 通过由3D卷积构成的低分辨率视频浅层特征提取网络 (LR video shallow feature extraction net, LFENet) 独立地提取视频帧的浅层特征, 然后快速空时残差块 (Fast spatio-temporal residual block, FRB) 仍利用 3D 卷积来建模视频的时空相关性. 为了减少 3D 卷积的计算开销, FRB 将原本一个大小为 $k \times k \times k$ 的卷积核解耦为两个连续的大小分别为 $1 \times k \times k$ 和 $k \times 1 \times 1$ 的卷积核. 这些模块生成的特征与全局残差学习 (Global residual learning) 模块提供的残差相融合, 最终经由上采样超分辨率网络 (LR feature fusion and up-sampling SR net) 生成高分辨率的重建视频.

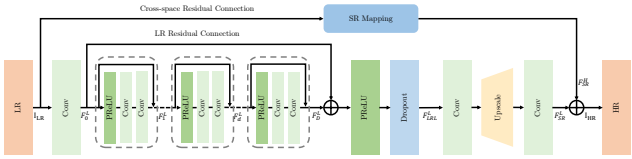

图 11　FSTRN 结构图

Fig. 11　Architecture of FSTRN

对于基于光流估计的运动补偿模式而言, 即使使用精确的光流信息, 图像层级的运动补偿也会在图像结构周围造成伪影, 并一直存在于重建的高分辨率的视频帧中. Tian 等[51]将可变形卷积用于帧间对齐, 预测相邻帧特征层面的偏移量来实现自适

应的帧间对齐, 而无需计算光流, 结构如图 12 所示. 该网络包括时序可变形对齐网络 (Temporally-deformable alignment network, TDAN) 和超分辨率重建网络. 其中, TDAN 由特征提取、可变形卷积和对齐帧重建三部分组成, 预测输入的多个低分辨率视频帧变形到参考帧的对齐结果. 这些对齐后的帧与参考帧一同经由超分辨率重建网络生成高分辨率的参考帧. 具体地, 输入网络的支持帧和参考帧经由堆叠的卷积层和残差块被映射到高维的特征空间, 进而通过卷积神经网络学习各个采样位置的偏移量输入可变形卷积. 基于学习到的偏移量, 原本形态固定的卷积核可以根据输入自适应地调整卷积操作过程中在特征图上的实际采样位置, 从而将支持帧的特征与参考帧对齐.

由于单一尺度不能很好地应对帧间较大的运动变化, Wang 等[67]提出了增强的可变形视频复原网络 (Enhanced deformable video restoration, EDVR), 结构如图 13 所示. EDVR 包括两个关键组件: 金字塔、级联、可变形对齐模块 (Pyramid, cascading and deformabel, PCD) 和时空注意力融合模块 (Temporal and spatial attention fusion, TSA), 分别用于应对视频中幅度较大的运动和多帧的有效融合. 具体地, EDVR 以局部时间窗口内的参考帧和 N 个支持帧作为输入, 金字塔和级联结构构成的 PCD 模块将所有支持帧在特征级别对齐到参考帧, 再采用基于注意力机制的 TSA 模块在特征空间中为与参考帧更相似的支持帧赋予更大的权重. 尽管 PCD 和 TSA 可以使网络的性能达到当时最先进的水平, 但其重建结果中出现了由输入帧模糊对运动补偿和细节融合造成的影响. 为此, EDVR 采用两阶段的复原策略, 即在主体网络后级联一个与 EDVR 相似但相对轻量的网络优化第一阶段的输出. 这一策略不仅消除了第一阶段中未解决的运动模糊问题, 而且减轻了重建帧间的不一致的现象.

无论是 3D 卷积还是可变形卷积, 隐式对齐存在没有有效融合相邻帧中的时序信息的问题. 以 3D 卷积为例, 相邻帧在通道维度级联后直接作为卷积层的输入, 但不同帧与参考帧的时间间隔并不会用作网络的先验信息, 导致无法有效融合信息. Isobe 等[69]设计的时间组注意力网络 (Temporal group attention, TGA) 根据支持帧到参考帧的时序距离将参考帧过去的 N 个相邻帧和未来的 N 个相邻帧划分为 N 个不同的时间组, 如图 14 所示, 这相当于为参考帧创建了 N 个不同帧率的数据. 例如, 对于给定的参考帧 I_t 及其支持帧 $\{I_{t-3}, I_{t-2}, I_{t-1}, I_{t+1}, I_{t+2}, I_{t+3}\}$, TGA 将其划分为三个时间组 $\{G_1, G_2, G_3\}$, 其中 $G_n = \{I_{t-n}, I_t, I_{t+n}\}$. 时间组通过由 2D 卷积和 3D 卷积堆叠而成的组内融合模块进行特征提取和组内时空信息的融合, 2D 卷积配备了由各组帧率确定的膨胀率. 随后, 融合后的特征由基于注意力的组间融合模块进一步优化后实现跨组的信息通信. 此外, TGA 结合了两帧间的单应性 (Homography) 和基于稳健性的退出机制, 设计了快速空间对齐模块, 提升了该模型在大运动的视频序列上的重建性能.

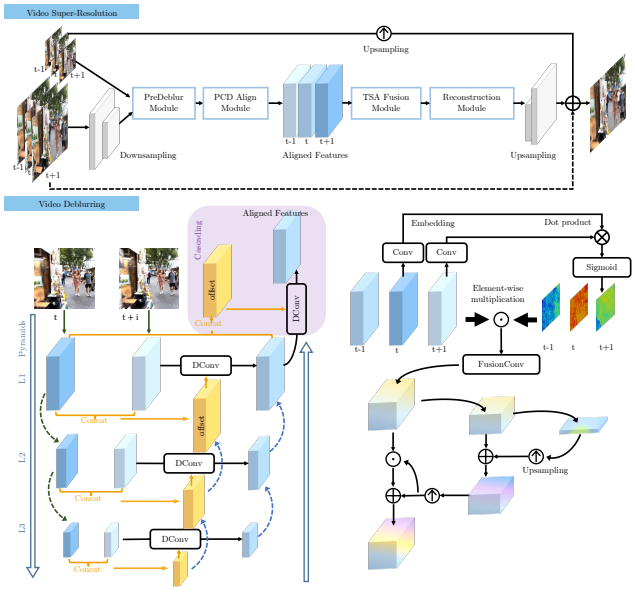

图 13 EDVR 结构图

Fig. 13 Architecture of EDVR

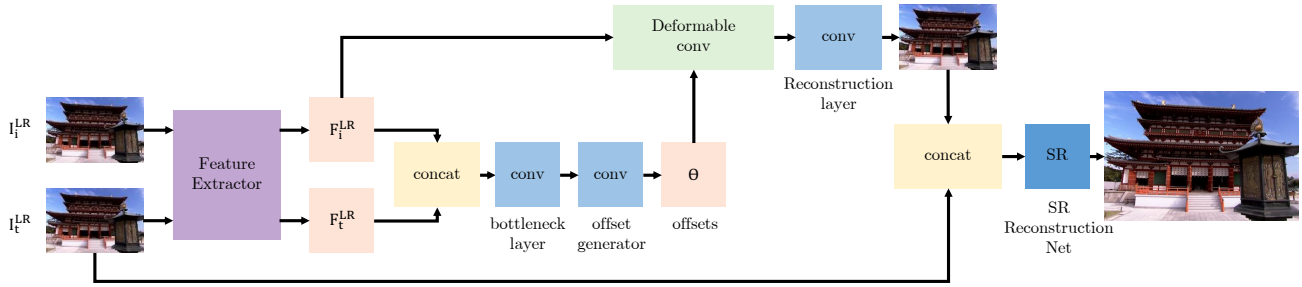

图 12 TDAN 结构图

Fig. 12 Architecture of TDAN

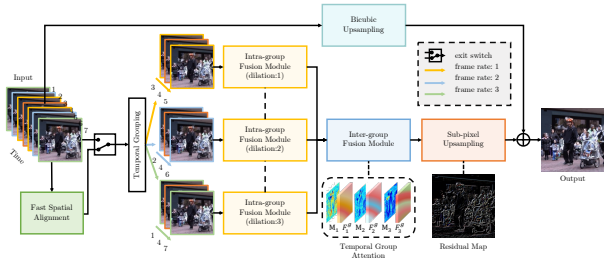

图 14　TGA 结构图

Fig. 14　Architecture of TGA

上述视频超分模型将对齐和回归过程独立建模, 忽视了相邻帧中共享相似内容和帧内不同位置可能包含相似结构的事实. Li 等[62]设计的多对应聚合网络 (Multi-correspondence aggregation network, MuCAN) 包含了一个时间多对应聚合模块 (Temporal multi-correspondence aggregation, TM-CAM) 和一个跨尺度非局部对应聚合模块 (Cross scale nonlocal correspondence aggregation, CN-CAM), 用于利用视频帧间和帧内的内容相似性, 结构如图 15 所示. 具体地, TM-CAM 将两个相邻的低分辨率视频帧编码为分辨率更低的特征后, 聚合单元 (Aggregation unit, AU) 从低分辨率特征空间开始逐层聚合, 从底层到高层的层级聚合方式可以实现大运动的补偿和精细的亚像素移动. 聚合单元采用基于块的匹配策略为参考帧的每一个图像块从支持帧中选取 K 个最相似的图像块, 用于自适应的上下文聚合. CN-CAM 在聚合的信息中进一步捕捉在帧内重复的非局部对应关系和空间模式. CN-CAM 利用平均池化操作构建了金字塔结构在不同尺度的特征图上完成非局部搜索, 并利用自注意力模块在特征融合前判断信息的有效性. 此外, MuCAN 还使用边缘感知损失来消除重建图像中的锯齿状边缘.

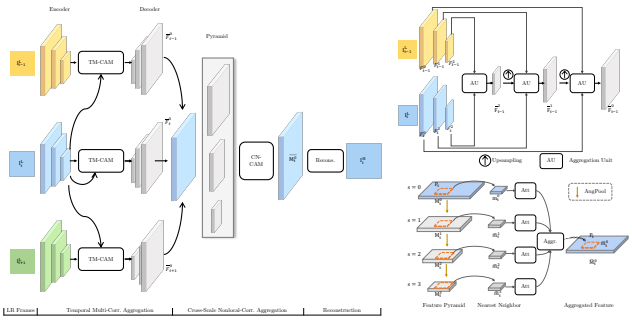

图 15　MuCAN 结构图

Fig. 15　Architecture of MuCAN

在不与参考帧对齐的情况下, Yu 等[108]设计了记忆增强的非局部注意力网络 (Memory-augmented non-local attention, MANA), 通过跨帧非局部注意力模块来保持帧间的时序一致性, 结构如图 16 所示. 具体地, 跨帧非局部注意力模块计算查询像素与键之间的相关性, 并以查询像素为中心的可训练高斯映射对相关性进行加权, 从而保持局部先验信息以减轻匹配错误的像素造成的影响. 为了解决相邻帧信息缺失的问题, MANA 引入了记忆增强注意力模块, 利用一个二维记忆库来存储在训练过程中学习到的训练集中其他视频中具有代表性的局部细节先验信息.

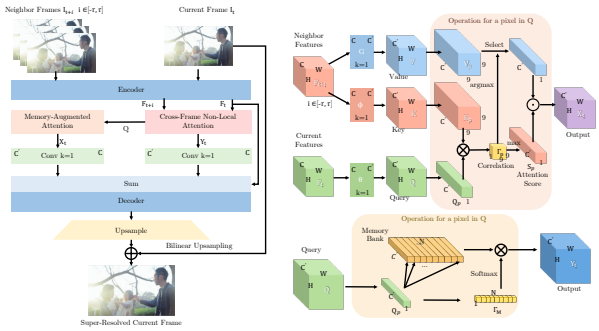

图 16　MANA 结构图

Fig. 16　Architecture of MANA

帧到帧的独立对齐方式因缺乏多个对齐过程的相关性而难以建模较长视频序列的运动信息, 而链式的渐进对齐方式则没有机会纠正先前传播过程中的对齐误差, 导致这些误差在传播过程不断累积放大. 因此, Zhou 等[84]重新思考了视频复原中的帧间对齐机制, 提出了迭代对齐模块 (Iterative alignment module, IAM), 通过逐步细化子对齐方式, 获得了更精确的运动补偿信息, 结构如图 17 (c) 所示. 具体地, IAM 在不同的长程对齐过程中将部分子对齐共享的特征作为先验知识, 迭代优化、矫正先前传播过程中的误差.

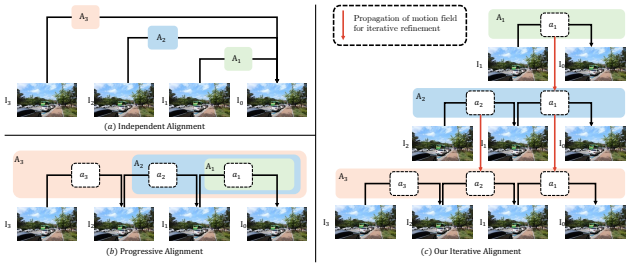

图 17　IAM 结构图

Fig. 17　Architecture of IAM

以上基于滑动窗口的 VSR 算法简单、有效, 具备良好的特征提取和帧间对齐性能. 它们通常将视频分成时间间隔固定且重叠的视频片段, 依次将每个片段或窗口中的视频帧输入 VSR 模型. 时间窗口包含了待重建的视频帧和少量来自过去以及未来的

视频帧信息, 以便 VSR 网络利用相邻帧的时空上下文信息更好地重建高分辨率的视频帧, 对具有显著运动或亮度变化的视频帧尤为有益. 重叠时间窗口划分也为输入视频片段的大小提供了一定的灵活性, 通过增加滑动窗口中的帧数可以提升 VSR 的性能, 但过多的帧数也会不可避免地增加计算开销, 反而限制了 VSR 的实际应用. 换言之, 由于时间窗口的重叠, VSR 网络不得不多次重复地处理同一视频帧, 显然会增加 VSR 模型的复杂度.

随着 Transformer 在自然语言处理领域中的突破性进展, 研究人员利用其并行计算能力解决 VSR 序列到序列的重复计算问题并逐渐取代了循环神经网络结构. 考虑到全连接自注意力层难以探索数据局部特性和词级前馈层缺乏特征对齐能力, Cao 等[71]首次将 Transformer 适配到 VSR 中, 设计了空-时卷积自注意力层和双向光流前馈层, 分别用于探索局部信息和挖掘相邻视频帧间的相关性, 结构如图 18 所示. 具体地, 该网络包括特征提取、Transformer 编码器以及重建网络三部分. 其中, Transformer 编码器包含了空-时卷积自注意力层和双向光流前馈层两个核心模块. 空-时卷积自注意力层采用三个独立的网络提取低分辨率视频帧的空域信息, 然后计算查询与键之间的注意力矩阵用于值的加权求和, 得到了增强的特征. 在自注意力操作中, 该网络计算所有时间块和空间块之间的相关性, 学习帧内和帧间的时空特征. 为了解决传统 Transformer 的全连接前馈层无法有效探索帧间相关性的问题, VSR Transformer 在前馈层中引入了由 SpyNet 估计得到的双向光流对齐视频帧.

为了减少 VSR Transformer 在时域和空域计算相似度的运算量, Liang 等[88]将 Swin Transformer 应用于单图超分的 SwinIR 扩展至视频超分任务, 设计了多尺度视频复原 Transformer (Video restoration transformer, VRT), 结构如图 19 所示. 具体地, VRT 将低分辨率视频划分为互不重叠的片段, 时序互自注意力模块 (Temporal mutual self attention, TMSA) 沿着时间维度进行片段间信息通信, 并聚合了视频序列蕴含的空时信息. TMSA 模块结合了基于滑动窗口的互注意力 (Mutual-attention) 和自注意力 (Self-attention), 分别用于对齐特征和提取时空特征. 与传统的自注意力相比, 互注意力的查询和键值分别来自不同帧 (如支持帧和参考帧), 得到的相似度描述了参考帧和支持帧中元素之间的相关性. TMSA 采用两次互注意力互相对齐两帧信息, 失去了原有的特征信息, 又引入自注意力补偿所丢失的信息. 空间维度上局部窗口的划分方式导致基于互注意力的帧间对齐无法很好地处理大运动视频. 因此, VRT 设计了平行变换 (Parallel warping) 机制, 采用基于光流引导的可变形卷积来进一步对齐视频帧, 并在四种不同的尺度上以不同大小的感受野完成特征提取, 在较小分辨率的视频特征上捕捉更大幅度的运动完成对齐, 而较大的分辨率则更有利于捕捉更多的局部相关性以聚合局部信息.

2.2 基于循环架构的 VSR 算法

循环神经网络擅长学习序列数据的非线性特征, 视频作为一种序列数据显然可以采用循环结构完成视频超分任务. 基于循环架构的视频超分辨率重建算法顺序地处理低分辨率的视频帧, 沿着时序维度不断更新和传播隐状态将来自较远时间的视频帧作为重建当前帧的参考信息. 与基于并行架构的 VSR 算法相比, 这类算法可以有效利用来自更大时间跨度的视频帧信息, 并通过对网络的跨帧复用以较为轻量的模型处理任意长度的序列, 却出现了长程信息丢失和难以分布式部署的问题. 按照信息传播方向的不同, 基于循环架构的算法可以划分为基于单向传播的算法和基于双向传播的算法. 其中, 基于单向传播的算法通常用于在线视频的超分辨率重建场景, 基于双向传播的算法通过建模过去和未来视频帧的相关性, 为当前帧的重建提供了更全面的信息, 可以有效提升视频超分辨率重建任务的性能.

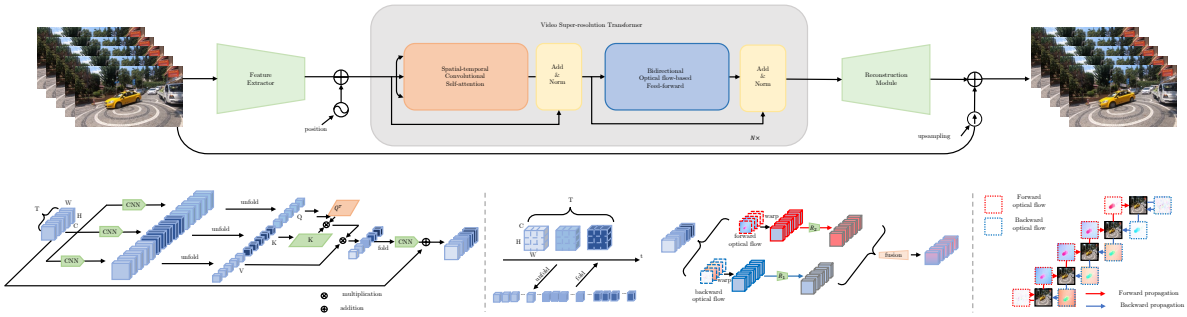

图 18 VSR Transformer 结构图

Fig. 18 Architecture of VSR Transformer

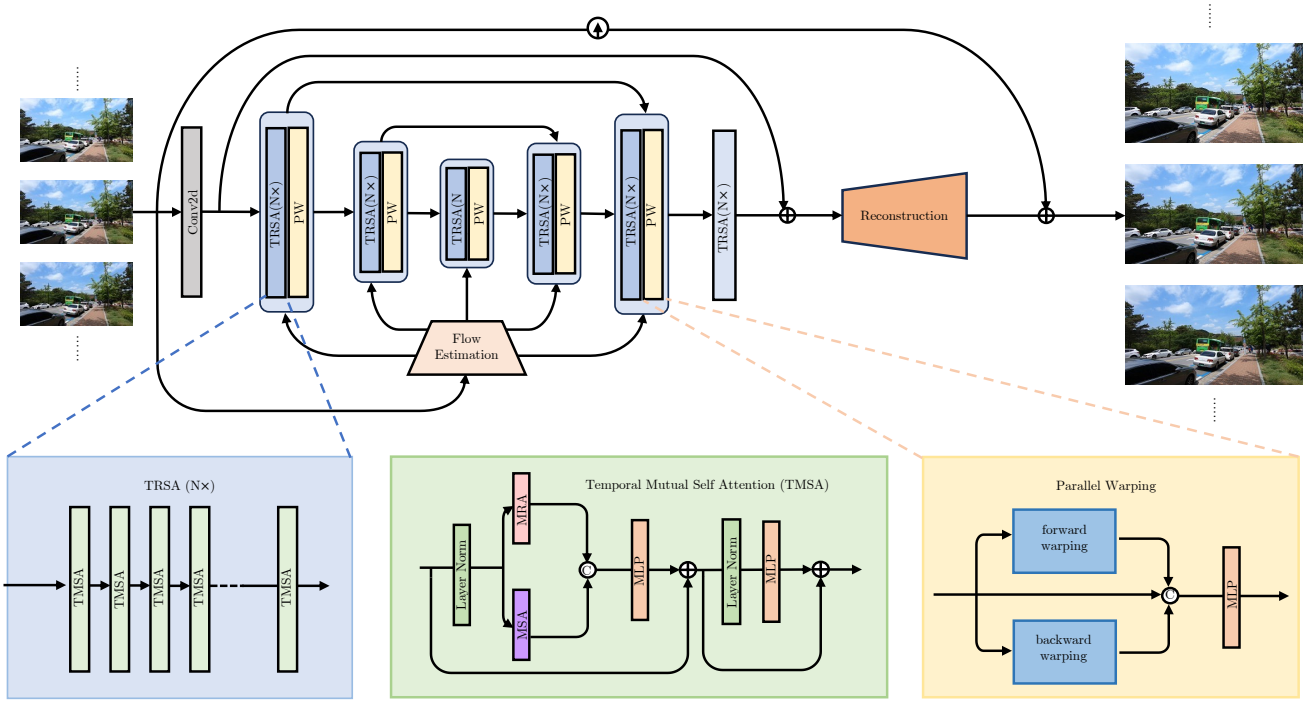

图 19　VRT 结构图

Fig. 19　Architecture of VRT

Tao 等[109]提出了揭示细节的深度视频超分辨率网络 (Detail-revealing deep video super-resolution, DRVSR), 利用 ConvLSTM 模块来处理时空信息, 结构如图 20 所示. 具体地, DRVSR 包含三个模块: 运动估计模块、基于亚像素运动补偿层 (Sub-pixel motion compensation, SPMC) 的运动补偿模块和融合模块. 其中, SPMC 可以根据预测的光流信息对相邻帧同时进行上采样和运动补偿操作, 包括网格生成器和采样器, 生成器先根据光流将低分辨率空间中的坐标转换到高分辨率空间, 采样器再在高分辨率空间中完成插值操作.

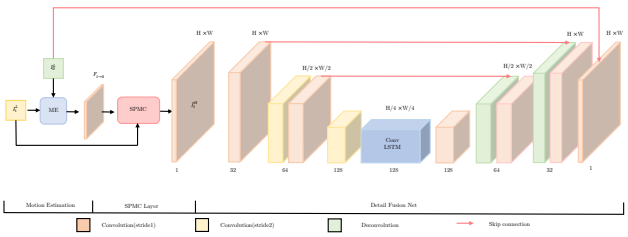

图 20　DRVSR 结构图

Fig. 20　Architecture of DRVSR

Sajjadi 等[102]设计的帧循环视频超分辨率重建网络 (Frame recurrent video super-resolution, FRVSR) , 利用已经重建的高分辨率帧迭代地估计后续帧以确保生成内容在时序上的一致性, 同时降低了网络的计算开销, 结构如图 21 所示. 具体地, FRVSR 先利用光流估计网络 FNet 计算低分辨率的前一帧到当前帧的光流信息, 再利用双线性插值将得到的低分辨率光流上采样至其高分辨率版本, 并与前一帧的高分辨率重建结果对齐后, 通过空间-深度变换操作获得对齐帧的低分辨率版本, 最后将对齐得到的高分辨率帧及其低分辨率版本输入超分重建网络获得目标帧的重建结果.

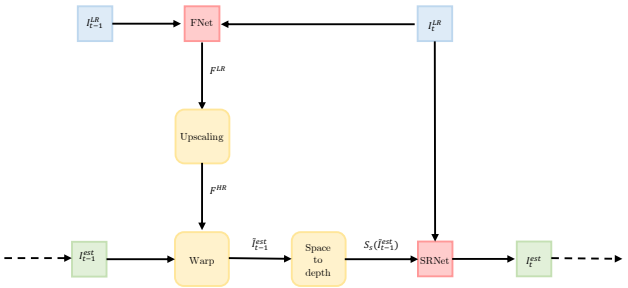

图 21　FRVSR 结构图

Fig. 21　Architecture of FRVSR

受 DBPN[117]的启发, Haris 等[59]设计了循环反投影网络 (Recurrent back-projection network, RBPN), 结构如图 22 所示. 具体地, RBPN 结合了循环神经网络与编码器-解码器结构, 用于聚合连续视频帧的时空上下文信息, 主要分为初始特征提取、多次投影和视频帧重建三个阶段. 其中, 初始特征提取阶段又分为单图超分辨率重建和多图超分辨率重

建两条路径, 前者直接将当前的低分辨率视频帧映射为特征张量 L, 后者合并当前帧与前一帧及两帧之间的光流信息并映射为特征张量 M. 多次投影模块聚合两路分支的特征, 通过迭代的放大和缩小过程优化目标帧的特征. 编码器-解码器结构以第 $k-1$ 个 L 特征和第 k 个 M 特征作为输入, 利用编码器将其投影为高分辨率的特征, 再由解码器下采样至低分辨率作为下一个多投影模块的输入, 最后将多个高分辨率特征合并后由重建模块生成高分辨率视频.

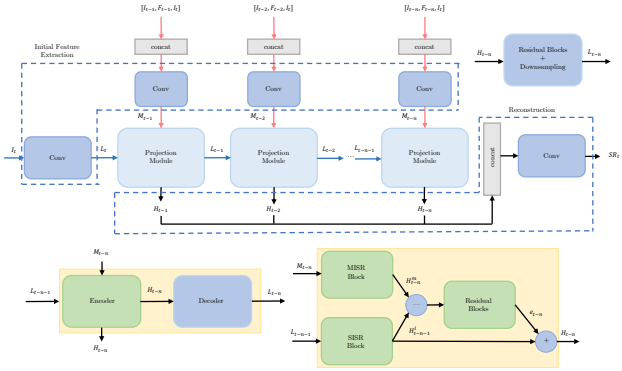

图 22 RBPN 结构图

Fig. 22 Architecture of RBPN

Fuoli 等[68]在循环潜在空间传播网络 (Recurrent latent space propagation, RLSP) 中将重建前一帧生成的隐状态引入当前帧的重建过程, 可以在设定的潜在空间中隐式地传播时序信息, 结构如图 23 所示. 具体地, RLSP 不需要显式的运动估计和运动补偿, 在通道维度级联当前帧与前后时刻的相邻帧后, 与前一时刻的重建帧得到当前帧的重建结果. 为了实现空间分辨率的对齐, RLSP 通过空间转通道的维度变换将高分辨率重建帧映射到低分辨率空间.

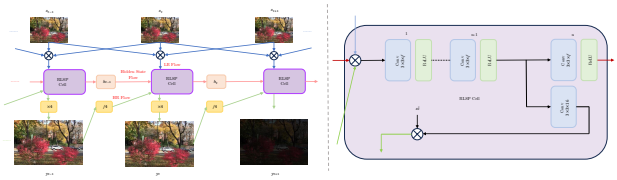

图 23 RLSP 结构图

Fig. 23 Architecture of RLSP

Isobe 等[103]提出的递归结构-细节网络 (Recurrent structure-detail network, RSDN) 是由双路结构-细节模块 (Two-stream structure-detail, SD) 组成的循环神经网络, 将每一帧分解为独立传播的结构和细节分量, 分别描述视频帧的轮廓信息和高频细节信息, 结构如图 24 所示. 具体地, 在传播过程中, 这两种信息在 SD 模块中不断交互, 可以在增强视频帧结构的同时有效恢复丢失的细节信息. 此外, RSDN 将隐状态视为历史字典, 利用隐状态适配模块构建相似性矩阵为隐状态施予不同的权重, 可以达到突出潜在的有用信息并抑制过时信息的目的.

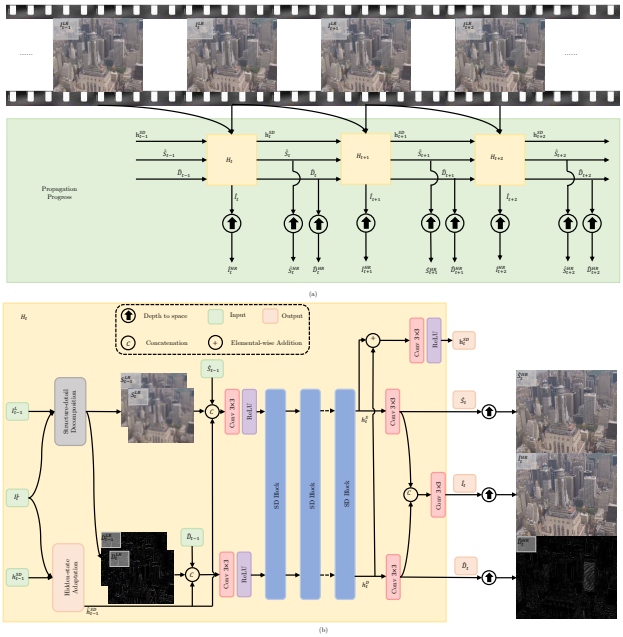

图 24 RSDN 结构图

Fig. 24 Architecture of RSDN

Isobe 等[101]对比了基于 2D CNN、慢融合的 3D CNN 和循环神经网络的视频超分辨率重建算法的运行时间. 实验结果表明, 3D 结构优于 2D 结构, 运行速度慢了约5~10倍, 循环神经网络重建的性能最好、参数量最少. 由此, Isobe 等[101]设计了循环残差网络 (Recurrent residual network, RRN), 在具有恒等跳连接的层之间引入了残差映射, 可以长期保存纹理信息并缓解了梯度消失的现象, 有助于更好地处理较长的视频序列, 结构如图 25 所示.

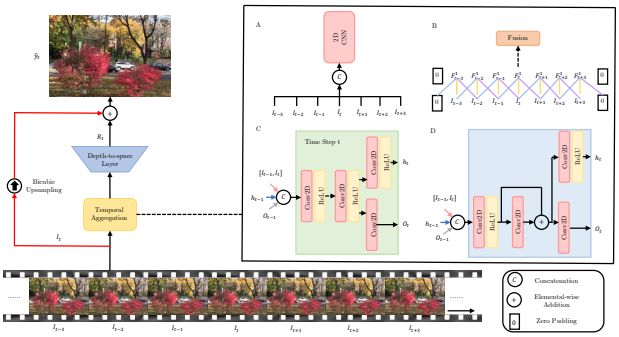

图 25 RRN 结构图

Fig. 25 Architecture of RRN

Fuoli 等[64]将可变形注意力金字塔 (Deformable attention pyramid, DAP) 引入基于循环

架构的视频超分辨率重建网络中, 以满足在线视频超分辨率重建场景的需求, 结构如图 26 所示. 具体地, DAP 由多尺度编码器、可变形注意力模块和迭代优化模块三个部分构成, 将单向循环网络中的隐状态与当前帧对齐. 为了降低传统注意力模块的计算开销, 可变形注意力模块动态地从特征图中选择有限数量的空间位置. DAP 将第 l 层编码器生成的当前帧的特征映射为查询, 键和值则交由网络动态地从前一帧的特征中采样 k 个像素点, 再由基于点乘和 SoftMax 的注意力计算以及逐层迭代优化实现帧间对齐.

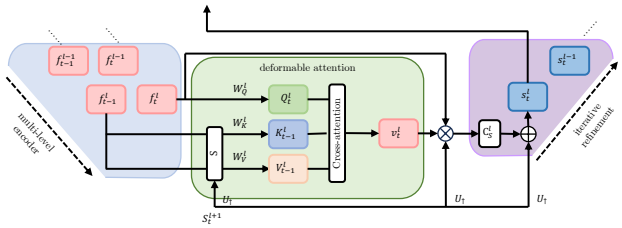

图 26　DAP 结构图

Fig. 26　Architecture of DAP

为了解决现有模型在时间建模方面的复杂度高和对遮挡或复杂运动场景下建模能力有限的问题, Isobe 等[82]构建了在低分辨率和高分辨率空间中显式地建模时序差异的模型 (Explicit temporal difference modeling, ETDM), 结构如图 27 所示. 具体地, 在低分辨率空间中, ETDM 计算参考帧与相邻帧之间的差异并利用区域分解模块 (Region decomposition module) 将其分解为低方差 (Low-variance, LV) 和高方差 (High-variance, HV) 区域, 分别对应整体变化较少和外形差异巨大的区域. ETDM 关注 LV 区域中相邻帧间的细节信息, 从 HV 获取粗略的补充信息, 分别利用具有不同大小感受野的卷积操作进行处理. 在高分辨率空间中, ETDM 预测相邻帧间超分辨率重建结果的差异, 使当前帧的重建结构能够从过去和未来帧的初步超分结果中获益.

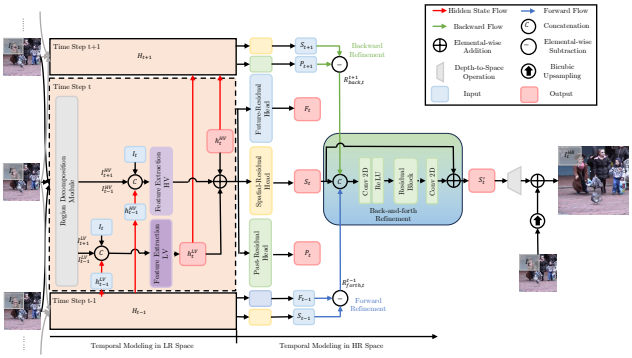

图 27　ETDM 结构图

Fig. 27　Architecture of ETDM

现有的在线视频超分算法大多是独立地对每一帧进行运动估计和运动补偿, 造成了计算冗余. 基于相邻帧的运动矢量高度相关的假设, Zhang 等[61]设计了在线视频超分的时序运动传播算法 (Temporal motion propagation, TMP), 代替复杂的运动估计方式加速帧间对齐过程, 结构如图 28 所示. 具体地, TMP 并非从零开始预测当前帧的运动矢量, 而是继承了前一帧的运动矢量并快速微调以适配当前帧. 同时, 考虑到视频内容中不同的运动类型, TMP 将运动解耦为对象运动和相机运动, 分别描述当前帧中对象的潜在运动状态和静态区域的位置变化情况, 独立传播两种运动类型为当前帧的每个像素生成了多个候选的偏移量, 再传递到相邻帧进行微调, 与距离最小的像素进行匹配.

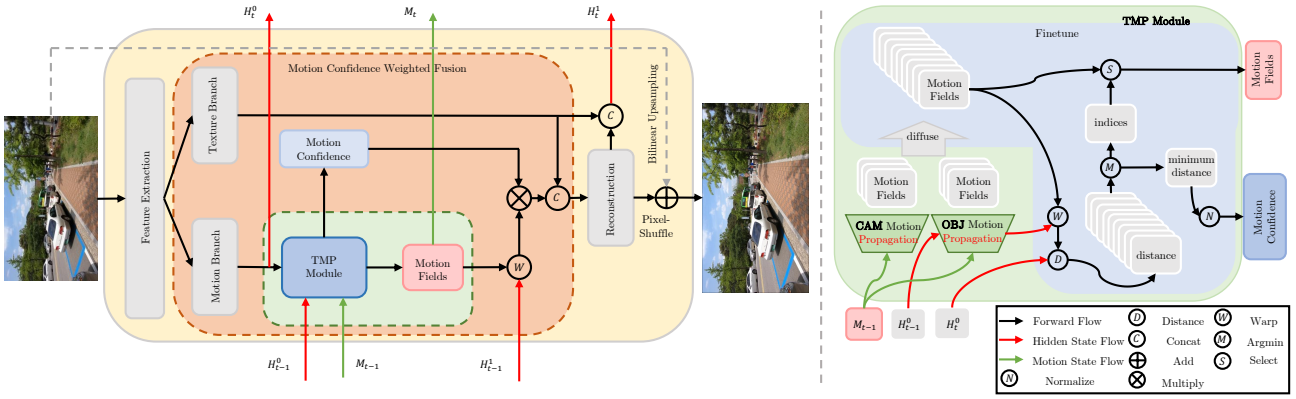

图 28　TMP 结构图

Fig. 28　Architecture of TMP

基于单向传播的算法只能访问过去和有限数量的未来帧, 难以捕捉视频序列中复杂的时序上下文关系. 此外, 可用时序上下文的缺失会导致视频序列中较早时刻重建帧质量较低的问题, 并在后续时刻随着时序信息的累积而逐渐改善. 基于双向传播的算法可以同时利用过去和未来视频帧的信息. 如 Huang 等[99]设计了基于双向传播的循环神经网络 (Bidirectional recurrent convolutional Network, BRCN) 建模视频序列的长程上下文信息, 结构如图 29 所示. BRCN 包括前馈和反馈两个子网络, 分别用于建模待重建帧与前、后帧特征之间的关联性. 前馈子网络包含输入层、输出层和两个隐藏层, 这些层的特征映射通过 3D 前馈卷积和循环卷积连接起来. 其中, 3D 前馈卷积用于连接当前时刻的输入层和隐藏层以及两个连续的隐藏层, 而循环卷积则连接了相邻帧之间的隐藏层. BRCN 通过循环神经网络建模了时序的长程依赖关系, 因未进行显式的运动估计和运动补偿, 导致无法达到最优的 VSR 性能.

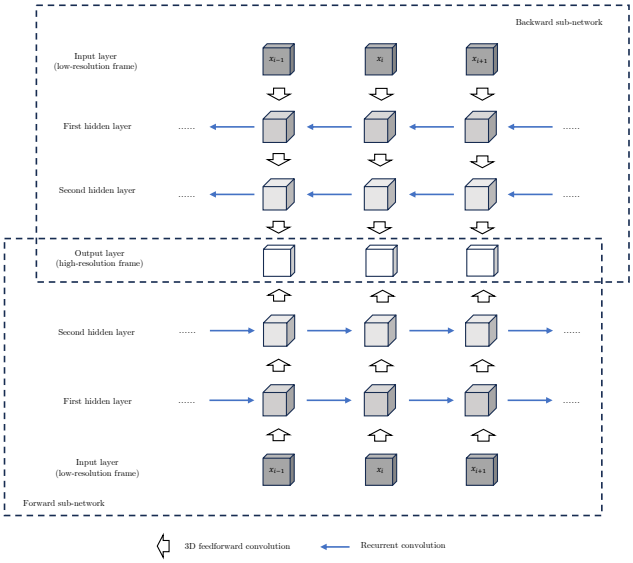

图 29 BRCN 结构图

Fig. 29 Architecture of BRCN

Li 等[44]设计了双向残差递归卷积网络 (Residual recurrent convolutional network, RRCN), 利用 CLG 变分法对目标帧及其相邻帧进行运动估计和运动补偿, 并将补偿后的视频帧输入递归网络获得正向残差和反向残差, 结构如图 30 所示. 为了重建目标帧中的细节信息, RRCN 会聚合所有的残差图. 在 RRCN 的基础上, RRCN+ 和 RRCN++ 引入了自集成 (Self-ensemble) 策略和单图像超分辨率网络 EDSR+, 从而获得了更佳的高分辨率视频的重建效果.

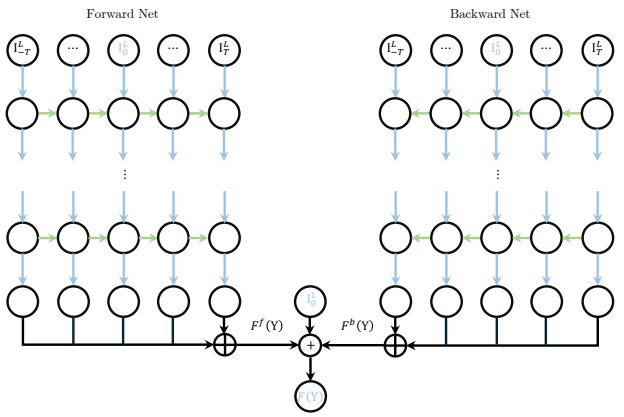

图 30 RRCN 结构图

Fig. 30 Architecture of RRCN

Yi 等[57]提出了渐进式融合非局部网络 (Progressive fusion non-local, PFNL), 该网络由一系列渐进式融合残差块 (Progressive fusion residual block, PFRB) 组成, 并引入非局部注意力来建模视频帧之间的时空相关性, 结构如图 31. PFNL 首先利用非局部残差块 (Non-local residual block, NLRB) 计算每个像素与所有帧中其他像素之间的相关性获得时空特征, 替代了复杂的运动估计和运动补偿过程. 接着, PFRB 经过多层卷积、融合这些特征获得包含独立空间信息和混合时间信息的特征.

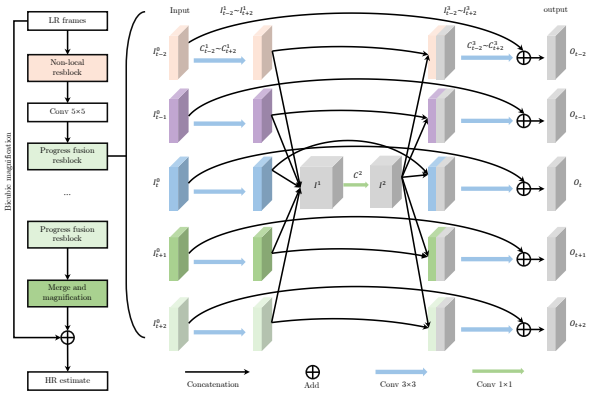

图 31 PFNL 结构图和PFRB 细节图

Fig. 31 Architecture of PFNL and Detail of PFRB

受可逆块的启发, Zhu 等[49]提出了残差可逆空时网络 (Residual invertible spatio-temporal network, RISTN), 使用残差可逆块 (Residual invertible block) 学习细粒度的特征表示以保持低分辨率帧与超分辨率重建帧之间的结构一致性, 结构如图 32 所示. 具体地, RISTN 包括空间模块、时间模块和重建模块三部分. 其中, 空间模块包含多个平行的残差可逆块, 用于提取作为时间模块输入的层次特征, 形成描述低分辨率和高分辨率视频帧之间差异的特征图; 时间模块利用具有残差密集卷积的长

短期记忆神经网络 (Residual dense convolutional LSTM, RDC-LSTM) 处理特征图, 以捕获连续视频帧的时间信息, 同时有效转换不同层次的空间特征; 重建模块利用稀疏特征融合的策略有选择地融合特征获得高分辨率的视频帧.

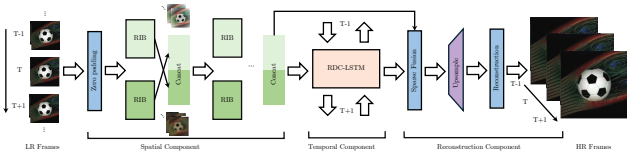

图 32　RISTN 结构图

Fig. 32　Architecture of RISTN

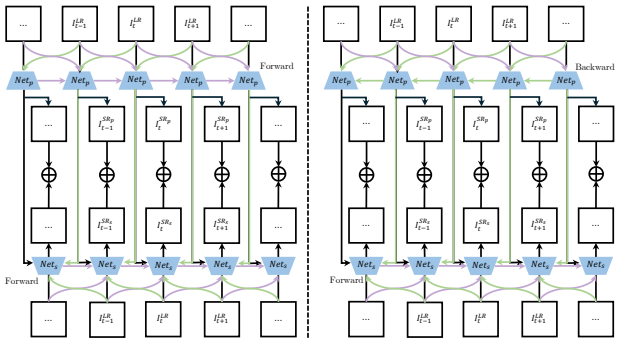

图 33　LOVSR (左) 和GOVSR (右) 结构图

Fig. 33　Architectures of LOVSR (left) and GOVSR (right)

Yi 等[106]提出了全能视频超分辨率重建网络 (Omniscient video super-resolution, OVSR), 结合了视频超分的滑动窗口和循环方法, 并将过去与未来的隐状态作为当前帧的重建信息, 结构如图 33 所示. OVSR 包含前置网络和后置网络, 前置网络从所有低分辨率视频帧中获得初始的超分辨率重建帧和隐状态, 后置网络在前述信息的辅助下重建所有低分辨率视频帧, 二者生成的超分结果相加作为最终的重建结果. 按照前置网络和后置网络的处理方向, OVSR 分为局部全能网络 (Local omniscient VSR, LOVSR) 和全局全能网络 (Global omniscient VSR, GOVSR). 其中, GOVSR 可以同时利用序列中的所有低分辨率图像进行超分辨率重建, 即前置网络和后置网络的处理方向相反; 当二者的处理方向相同时, LOVSR 仅能利用过去和有限的未来帧信息. 换言之, LOVSR 和 GOVSR 分别适合应用在线和离线视频超分场景.

基于深度学习的视频超分通过利用时序信息提升其超分性能, 却也因引入相应的处理模块导致网络复杂难以复用. Chan 等[75]梳理视频超分的框架和传播、对齐、聚合及上采样四个组件, 在对现有方案复用的基础上经过细微改动后, 提出了简单而有效的视频超分辨率重建方案 BasicVSR, 结构如图 34. 不同于时序信息在局部窗口内或整个视频序列上的单向传播方式, BasicVSR 采用双向循环神经网络, 可以充分利用整个视频信息, 解决了不同帧信息接收不均衡的问题. 在视频帧对齐方面, BasicVSR 利用光流网络预测相邻帧间的光流信息, 在特征层面实现帧间对齐, 解决了不精确的光流估计引起图像层面对齐出现的模糊和伪影问题. 在 BasicVSR 的基础上, Chan等[75]引入了信息重填 (Information-Refill) 和耦合传播 (Coupled Propagation) 两种机制消除传播过程中的累积误差, 提出了性能更强的视频超分网络 ICONVSR.

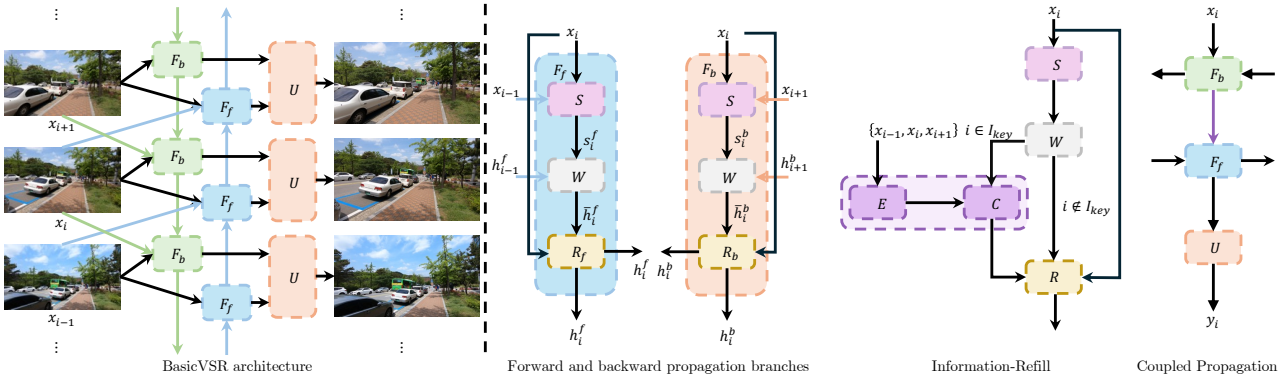

图 34　BasicVSR (左) 和ICONVSR (右) 结构图

Fig. 34　Architectures of BasicVSR (left) and ICONVSR (right)

Liu 等[87]在 BasicVSR 架构的基础上设计了轨迹感知 Transformer (Trajectory-aware transformer, TTVSR), 进一步探索视频中更有效的时空信息建模机制, 结构如图 35 所示. TTVSR 以视频中的帧为视觉特征, 将时空中连续的视觉特征定义为一系列在内容上预对齐的运动轨迹. 自注意力机

制只沿着同一运动轨迹查询每个视觉特征. 为了实现这一操作, TTVSR 巧妙地设计了位置图机制, 对预先定义的视觉特征的坐标位置图进行运动变换达到建模视觉特征轨迹的目的. 对比在整个时空维度进行自注意力计算, TTVSR 有效降低了计算开销, 提高了模型建模长距离视频特征的能力.

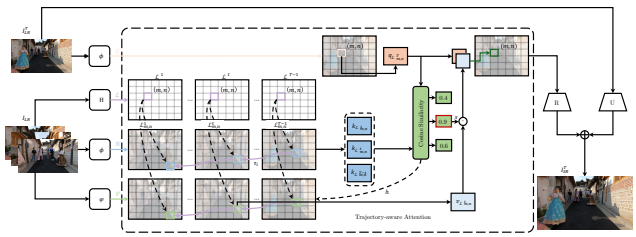

图 35 TTVSR 结构图

Fig. 35 Architecture of TTVSR

因未考虑空间信息的重要性, TTVSR 对短视频序列的处理效果不佳. 为此, Tang 等[89]联合用于空间特征提取的空间增强网络 (Spatial enhanced network, SEN) 和时序信息对齐的时序轨迹增强网络 (Temporal-trajectory enhanced network, TEM), 设计了用于视频超分辨率重建的协作 Transformer (Collaborative transformer, CTVSR), 结构如图 36 所示. 具体地, SEN 利用令牌丢弃注意力 (Token dropout attention, TDA) 和可变形多头交叉注意力 (Deformable multi-head cross attention, DMCA) 来建模局部和全局的空间信息, 代替了 TTVSR 中的卷积操作. 不同于 TTVSR 仅选择同一运动轨迹中最相似的图像块, TEM 筛选了一定数量的参考图像块与当前图像块对齐, 减轻了有限时序上下文和单一图像块中伪影对网络性能的影响.

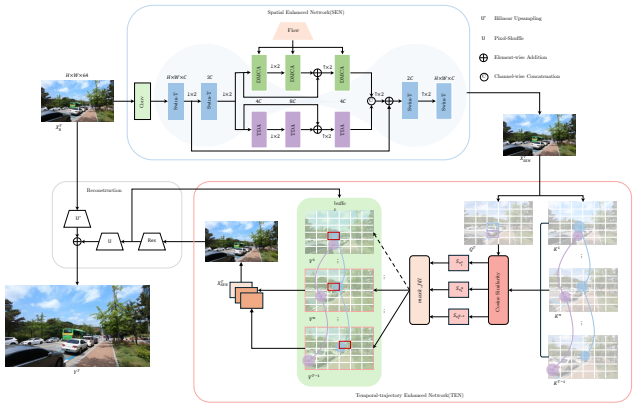

图 36 CTVSR 结构图

Fig. 36 Architecture of CTVSR

Lee 等[29]首次提出了基于参考的视频超分辨率重建网络 (Reference-based video super-resolution, RefVSR), 结构如图 37 所示. 具体地, RefVSR 利用手机中不同焦距的镜头同时捕获低分辨率视频和具有更多细节的高分辨率视频, 将后者作为前者的辅助信息提高 VSR 的性能. 在 BasicVSR 双向分支的循环单元中, 上一时刻的隐状态不仅包含了过去或未来低分辨率帧中的信息, 而且汇聚了过去或未来高分辨率参考视频帧的细节, 因此参考帧的时序信息在帧间对齐的过程中随之传播. 为了利用当前时刻的参考视频帧, RefVSR 设计了参考对齐和传播模块, 主要包含余弦相似度、参考对齐和传播时序融合三个子模块. 其中, 余弦相似度模块利用高分辨率参考视频帧和待重建低分辨率帧间的余弦相似度矩阵为其他两个模块提供了索引图和置信图; 参考对齐模块利用索引图将从参考帧提取的特征对齐到低分辨率视频帧, 传播时序融合模块将其与之前聚合了帧间信息的特征进行融合.

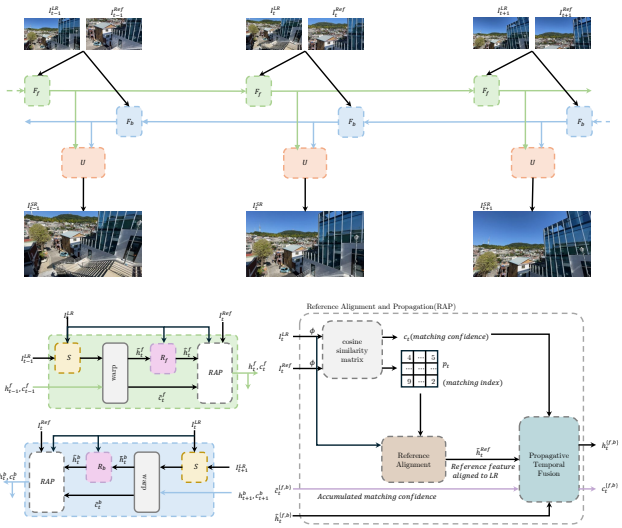

图 37 RefVSR 结构图

Fig. 37 Architecture of RefVSR

与 RefVSR 不同, Jiang 等[81]提出的 C^2-Mathching, 没有参考高分辨率视频信息, 仅仅利用从相似场景中拍摄的一帧高分辨率图像作为额辅助信息, 结构如图 38 所示. 为了解决视频帧和参考图像在形变和分辨率两方面的差异, C^2-Mathching 分别设计了对比性对应网络 (Contrastive correspondence network) 和教师-学生关联蒸馏法 (Teacher-student correlation distillation). 具体地, 对比性对应网络利用三元组损失函数 (Triplet margain loss) 学习低分辨率帧和高分辨率参考图像间鲁棒形变的对应关系, 再用教师-学生关联蒸馏法进一步提升对比性对应网络建模纹理相关性的准确度. 教师网络建模了高分辨率-高分辨率图像间的纹理对应关系, 并利用知识蒸馏的策略指导低分辨率-高分辨率图像的纹理匹配.

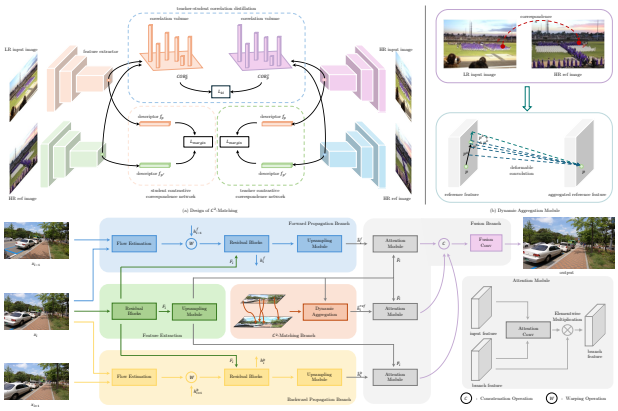

图 38　C^2-Mathching 结构图

Fig. 38　Architecture of C^2-Mathching

尽管长程传播在非盲设定下取得了成功, 但在面对真实场景中复杂多样的退化时, 表现却不尽如人意. Chan 等[28]将 BasicVSR 拓展到真实场景, 提出了 RealBasicVSR 网络, 结构如图 39 所示. 为了解决长程传播过程中退化方式复杂和噪声被放大的问题, RealBasicVSR 引入了一个动态优化模块, 将低质量视频帧自适应地进行多次预处理后输入到 BasicVSR. 实验结果表明, RealBasicVSR 采用更长的序列训练 VSR 网络可以更好地利用长程信息, 这比使用更大的训练批次 (Batch) 以获得稳定的权重下降梯度更为重要.

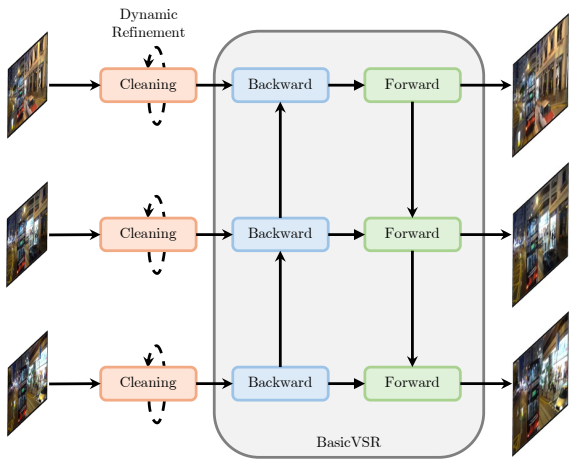

图 39　RealBasicVSR 结构图

Fig. 39　Architecture of RealBasicVSR

类似地, Qiu 等[80]针对存在噪声、压缩伪影的视频, 设计了退化鲁棒的空-时-频 Transformer (Spatiotemporal frequency-transformer, FTVSR), 结构如图 40 所示. 具体地, 在空间-时间-频率的复合域中, FTVSR 将视频划分为图像块, 再转换为频谱图, 频谱图的每一个通道代表了一个频率. 这种表示形式有助于 FTVSR 沿着频率带计算自注意力, 更有效地区分真实纹理和压缩伪影. 同时, 为了应对视频的复杂退化过程, FTVSR 采用了双频率注意力机制来建模全局和局部的频率关系, 分别处理依赖于不同空间范围的噪声和压缩伪影等退化因素. 此外, FTVSR 发现在空间-时间-频率复合域中依次执行空间-频率注意力计算和时间-频率注意力计算是最有效的模式, 这是因为消除空间上的伪影有益于建立时序关系.

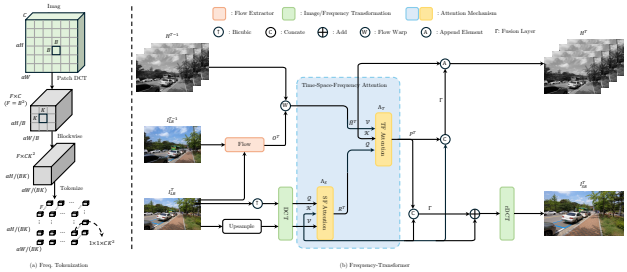

图 40　FTVSR 结构图

Fig. 40　Architecture of FTVSR

针对 BasicVSR 在时序信息传播和帧间对齐方面的不足, Chan 等[83]引入二阶网格传播和光流引导的可变形对齐提出了 BasicVSR++, 结构如图 41 所示. 在信息传播的过程中, BasicVSR++ 首先引入网格传播建立前向和后向分支间的信息交换, 利用二阶马尔科夫性质改进了二阶网格传播, 即当前帧不仅接受来自相邻帧的隐状态, 还接受来自更远一帧的隐状态. 这种设计方式有助于在当前帧的重建过程中引入更多的时序信息, 可以有效解决遮挡和边界问题. 同时, 为了解决可变形卷积训练不稳定造成的偏移量溢出问题, BasicVSR++ 使用光流指导可变形卷积学习偏移量, 不仅可以获得丰富的可变形卷积偏移量, 还可以提升网络训练的稳定性. BasicVSR++ 也因其相对轻量的网络规格和卓越的重建性能而成为后续研究工作的基石.

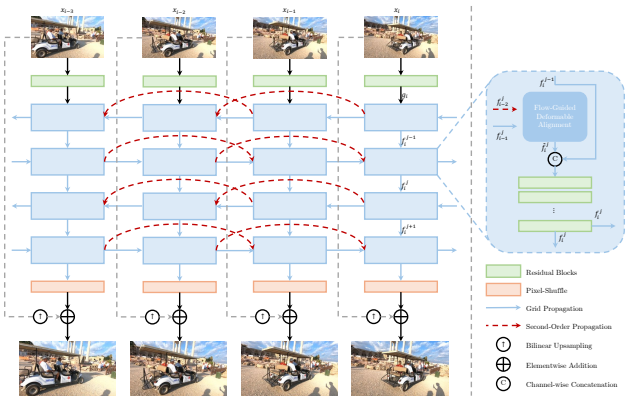

图 41　BasicVSR++ 结构图

Fig. 41　Architecture of BasicVSR++

例如, 基于 BasicVSR++ 的网络结构和自注意力机制的特征提取模块, Shi 等[72]探究了现有帧间对齐方案在视频超分辨率重建 Transformer 的性能, 认为 Transformer 处理运动幅度较小的帧间变换效果好, 而现有帧间对齐方案会降低其性能. 这主要是因为不准确的帧间运动估计信息和帧间对齐过程中插值操作破坏了亚像素信息. 同样, Transformer 可以通过增大注意力窗口应对更大幅度运动的帧间运动, 这必然会带来额外的计算开销. 为此, Shi 等[72]设计了基于图像块的帧间对齐算法, 结构如图 42 所示. 该算法根据 Transformer 的窗口大小对图像分块, 计算图像块内的光流平均值估计块的跨帧对应位置, 降低光流估计不准确对超分性能的影响. 得到运动信息后, 该算法对整体图像块进行移动, 可以保留像素的相对位置关系, 使对齐结果更加锐利.

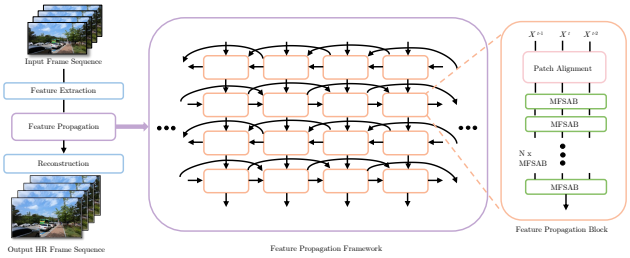

图 42 PSRT 结构图

Fig. 42 Architecture of PSRT

Xu 等[98]认为双线性插值等价于低通滤波, 会损失相邻帧中的高频信息, 最近邻插值则造成了高频信息的失真, 提出了隐式对齐的视频复原 Transformer (Implicit alignment restoration transformer, IART) , 在帧间对齐过程中将插值操作替换为计算局部窗口内的注意力, 结构如图 43 所示. 即使估计的光流存在误差, IART 也能达到良好的鲁棒性.

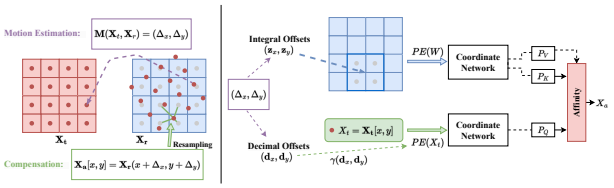

图 43 IART 结构图

Fig. 43 Architecture of IART

有限的感受野阻碍了卷积神经网络对长程时空依赖关系的建模, 而 Transformer 通过堆叠自注意力层获得了建模长程依赖能力却带来了巨大的计算开销. 为更好地兼顾模型计算效率和特征表示能力, Li 等[95]设计了在频域聚合时空信息的多频率表示增强模块 (Multi-frequency representation enhancement, MFE), 结构如图 44 所示. 相较于空域, 频域处理方式可以建模全局的特征, 却无法捕捉未对齐多张视频帧所蕴含的复杂时空信息.

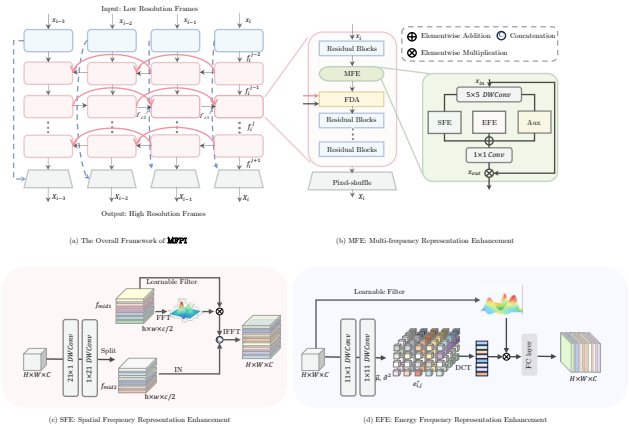

图 44 MFPI 结构图

Fig. 44 Architecture of MFPI

因此, MFE 设计了三个处理频谱图的分支, 即空间-频率表示增强分支 (Spatial-frequency enhancement branch, SFE) 、能量-频率表示增强分支 (Energy-frequency representation enhancement branch, EFE) 和一系列卷积层. 其中, SFE 利用快速傅立叶变换和大尺寸的卷积核来捕捉频域空间的长程依赖关系, EFE 借助设计的能量余弦变换来优化特征并进一步挖掘潜在的频率分量, 而卷积层也利用大的卷积核在空间域获取全局特征. 为了进一步提升模型的性能, MFE 针对输入数据的特点设计了特权信息训练策略, 通过编码高分辨率视频中的特权信息初始化模型参数和加快训练过程.

Dong 等[94]重新思考了视频超分中的频域表示和目标运动之间的关系, 提出了方向性的频域视频超分辨率重建网络 (Directional frequency video super-resolution, DFVSR), 结构如图 45 所示. 该网络利用方向性的频率表示方法 (Directional frequency representation, DFR) 同时描述图像的细节信息和结构信息以及物体在水平、竖直和对角线等不同方向的运动信息. 基于这一表示形式, DFVSR 在方向性的频率增强对齐模块 (Directional frequency-enhanced alignment, DFEA) 中利用任务相关信息进行双重增强, 有助于帧间对齐操作更加关注保真度高的区域, 既保留了有效信息、减少了无效信息, 也提高了可变形卷积训练的稳定性.

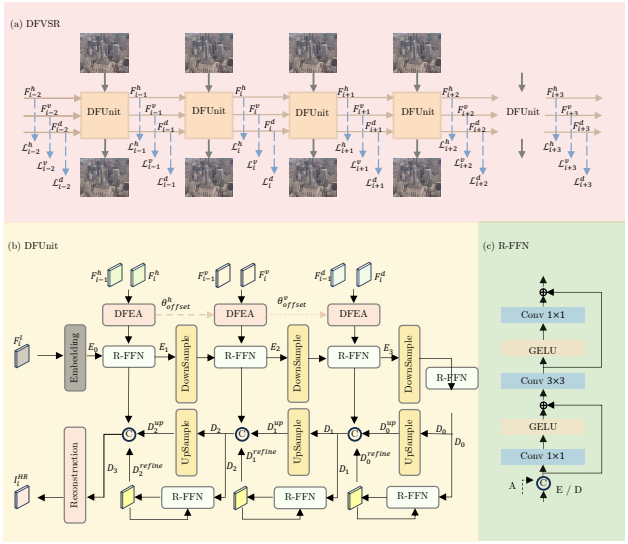

图 45　DFVSR 结构图

Fig. 45　Architecture of DFVSR

Zhou 等[97]在特征层面划分信息, 设计了掩码区分的帧内-帧间注意力机制 (Masked intra and inter frame attention, MIA-VSR), 以解决 Transformer 在计算资源有限的设备部署难的问题, 结构如图 46 所示. 不同于在通道维度直接级联之前的隐状态与当前帧输入注意力块, MIA-VSR 充分考虑了之前隐状态和当前帧的相关性从而更加合理地利用过去已增强的信息. 具体地, 在帧间-帧内注意力模块 (Inter-frame and intra-frame attention, IIA) 仅将当前帧的特征映射为查询张量 (Query), 之前的隐状态作为键 (Key) /值 (Value) 张量仅需为增强当前帧提供补充信息, 大幅降低了计算开销. 为了进一步消除计算冗余, MIA-VSR 根据相邻帧的相似性和连续性设计了自适应的掩码处理机制, 计算相邻帧特征层之间的差异后交由网络生成二进制掩码, 只允许少量的查询张量参与计算注意力.

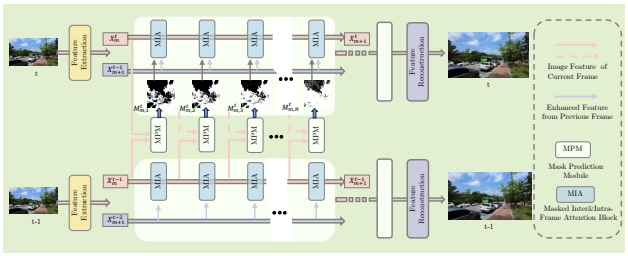

图 46　MIA-VSR 结构图

Fig. 46　Architecture of MIA-VSR

基于循环架构和基于 Transformer 的超分算法持续刷新着视频超分辨率重建任务的性能, 导致后续视频超分算法局限在这两类方法中. 基于循环架构的 VSR 算法采用参数共享方式减少了模型的参数量, 但依次处理视频帧的方式导致其难以并行部署. 该网络存在的梯度爆炸或消失、信息衰减、噪声放大等问题也导致其难以捕捉长程依赖信息. 而基于并行架构的 VSR 算法具有较好的并行化处理和部署能力, 模型参数和计算开销却非常巨大.

为了充分利用二者的优势, Liang 等[93]设计了循环视频复原 Transformer (Recurrent video transformer, RVRT), 结构如图 47 所示. 具体地, RVRT 将长度为 T 的视频序列划分为互不重叠的视频片段, 每个片段包含 N 个视频帧, 由 Transformer 并行化处理. 在处理当前视频片段时, RVRT 利用引导的可变形注意力 (Guided deformable attention, GDA) 直接将先前片段的特征与当前片段对齐而非帧到帧的对齐, 可以在片段之间以更大的隐状态传播时序信息. 可以看出, RVRT 使用并行结构提取局部特征, 利用循环架构提取全局特征, 达到了模型参数量、计算效率和性能之间的平衡.

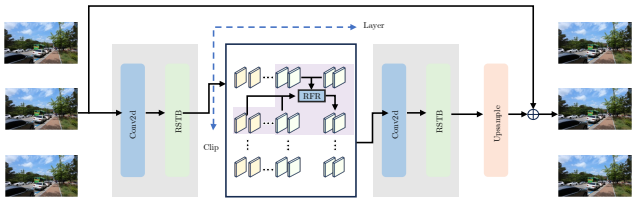

图 47　RVRT 结构图

Fig. 47　Architecture of RVRT

上述基于回归的方法主要使用诸如均方差的损失函数优化模型, 获得更高的基于向量范数的评价指标 (例如峰值信噪比), 重建的视频帧却缺乏更逼真的细节. 对抗训练方式可以在一定程度上改善单帧图像的视觉质量, 却难以兼顾视频序列中帧间的连贯性. Chu 等[100]首次提出了面向视频超分的时序一致性生成对抗网络 (Temporally coherent GAN, TecoGAN). 除了基于循环架构的生成器之外, TecoGAN 引入了结合运动补偿的时空判别器, 以连续三帧沿通道维度级联作为输入, 保证了重建图像的纹理真实感和连续帧的时序一致性, 结构如图 48 所示. 此外, TecoGAN 设计了一个时空对抗损失 (Ping-pong loss) 消除循环架构中的漂移伪影, 进一步提升了高分辨率视频帧的视觉质量.

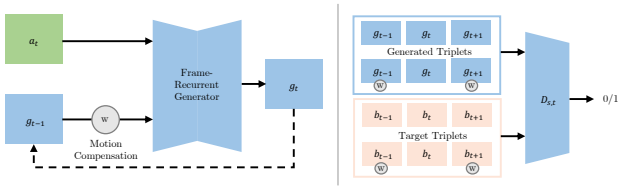

图 48　TeoGAN 结构图

Fig. 48　Architecture of TeoGAN

随着去噪扩散概率模型 (Denoising diffusion probabilistic model, DDPM) 出现, 基于扩散范式的生成模型在图像和视频生成任务上取得了巨大的成功. 许多工作将扩散模型用于视频超分辨率重建任务, 以生成纹理更逼真的视频. 如 Rota 等[55]通过引入时序条件模块 (Temporal conditioning module, TCM), 将在单图超分任务中预训练的扩散模型拓展至视频超分任务, 设计了视频超分潜在扩散模型 (StableVSR), 结构如图 49 所示. 具体地, TCM 利用光流对齐得到的时序纹理信息, 可以引导当前帧的生成过程沿着高感知质量和时序一致的方向进行. TCM 利用前一帧采样得到的重建结果和两帧之间的运动信息为生成当前帧提供丰富的纹理信息, 避免了中间采样过程引入的噪声干扰时序纹理. 与现有方法类似, StableVSR 采用了帧级别的双向采样策略, 充分利用从过去到未来的时序信息, 显著提升了重建视频的视觉质量.

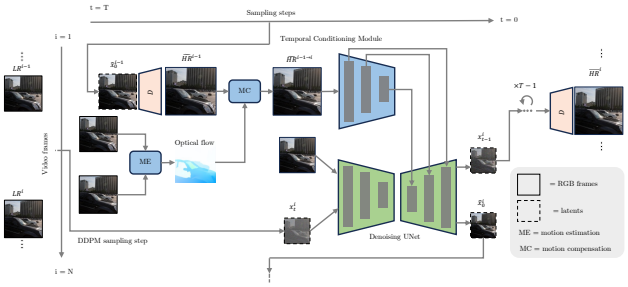

图 49 StableVSR 结构图

Fig. 49 Architecture of StableVSR

为了减少扩散过程内在的随机性对视频帧间时序一致性的影响, Yang 等[110]设计了面向真实世界运动引导的潜在扩散模型 (Motion-guided latent diffusion, MGLD), 在扩散采样过程中引入低分辨率视频序列的运动信息, 有效增强了视频帧在潜在特征空间的时序一致性, 结构如图 50 所示. 具体地, MGLD 首先将前向和后向光流下采样至潜在特征的维度, 然后利用运动信息将潜在特征对齐到相邻帧, 沿两个方向计算累积对齐误差. 遮挡的存在对光流估计造成了消极的影响, 干扰了采样过程导致最终采样结果出现了伪影. 为此, MGLD 网络预测了描述遮挡的掩码来矫正运动信息, 引导最后一步采样的潜在向量被送入 VAE 的解码器, 以获得最终的重建视频帧. 虽然运动引导的潜在扩散模型可以生成时序较为一致的视频内容, 但低分辨率潜在空间和高分辨率帧空间的域差距仍会造成解码器重构的细节在时间上不连贯. 因此, MGLD 在预训练好的 VAE 解码器上添加了沿时间维度的一维卷积, 构成了时序感知序列解码器, 以较高的连续性还原视频的细节信息.

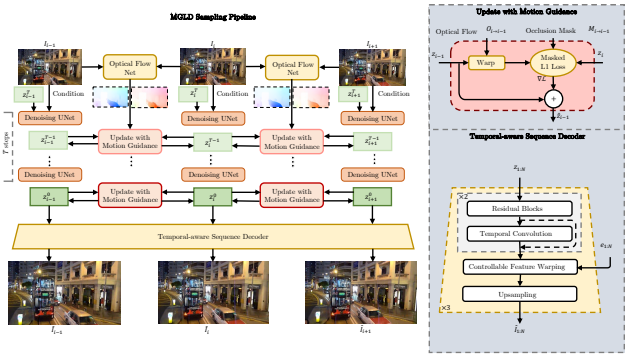

图 50 MGLD 结构图

Fig. 50 Architecture of MGLD

利用 3D 卷积或时间注意力将图像扩散模型迁移至视频任务的方法在一定程度上提高了生成视频的质量和稳定性, 但在潜在空间引导帧间对齐的方式难以约束底层的一致性, 纹理闪烁等问题依然存在, 且难以确保长视频序列中全局时间的一致性. 为此, Zhou 等[32]结合维持局部和全局时序一致性策略设计了视频超分潜在扩散模型 Upscale-A-Video, 结构如图 51 所示. 具体地, Upscale-A-Video 在局部向 U-Net 中插入 3D 卷积和时序自注意力层, 增强其对时间维度的建模能力, 使得 U-Net 可以学习视频序列中帧间的依赖关系, 保持了局部的时序一致性. 与此同时, VAE 解码器也在集成 3D 卷积和空间特征变换层 (Spatial feature transform, SFT) 后进行了微调, 以进一步提升视频的底层一致性, 减少纹理闪烁和色彩偏移. 考虑到 U-Net 作用范围有限难以约束长视频序列的全局一致性的问题, Upscale-A-Video 设计了光流引导的循环传播模块. 在不增加训练参数的前提下, 该模型利用估计的光流不断向过去或未来传播潜在特征, 有效延拓了模型的时序感知范围. 此外, Upscale-A-Video 还允许文本作为条件来指导模型生成更真实的细节.

3 视频超分中的帧间对齐方式

如何有效利用帧间信息是 VSR 过程中的重要环节. 帧间对齐是利用相邻帧间的相关性, 将相邻的支持帧或特征变形到待重建参考帧的过程. 现有的帧间对齐方法主要可以分为: 显式对齐法、隐式对齐法、混合对齐法和无需对齐法, 如图 52 所示.

显式对齐法是基于运动估计和运动补偿操作, 在视频超分辨率重建算法中广泛使用. 其中, 运动估计旨在提取帧间的运动信息, 运动补偿则利用运动信息在视频帧或特征级别进行变形操作以实现帧与帧的对齐. 如图 53 所示, 运动估计大多采用光流法, 假设在一帧中观察到的强度模式沿着时间维度保持不变, 任何变化都是由物体运动或相机移动造成的.

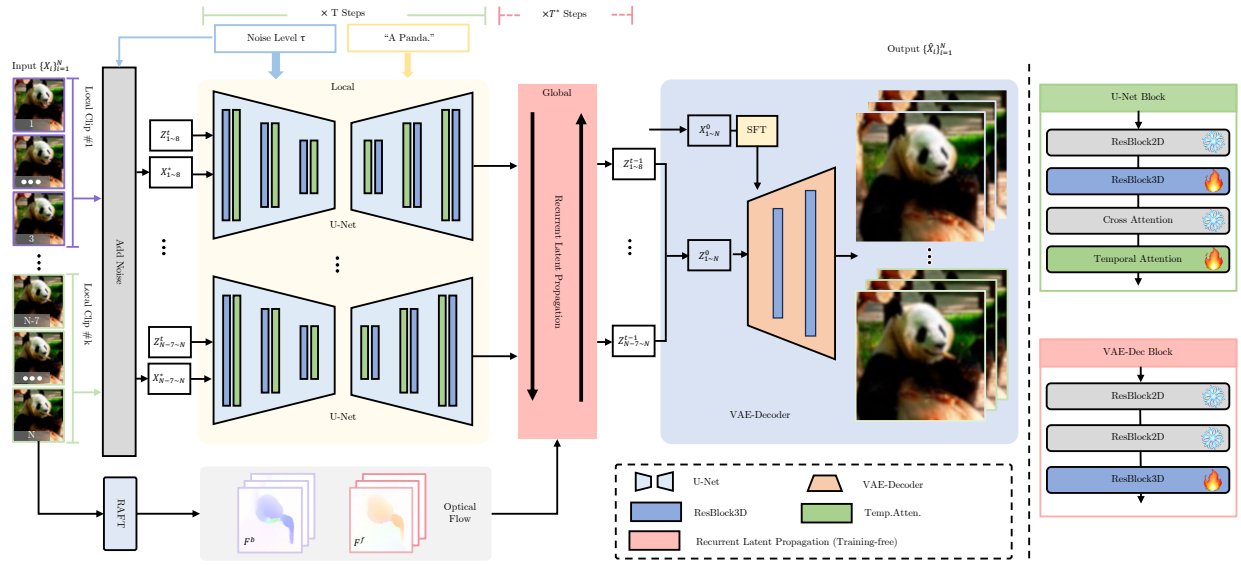

图 51　Upscale-A-Video 结构图

Fig. 51　Architecture of Upscale-A-Video

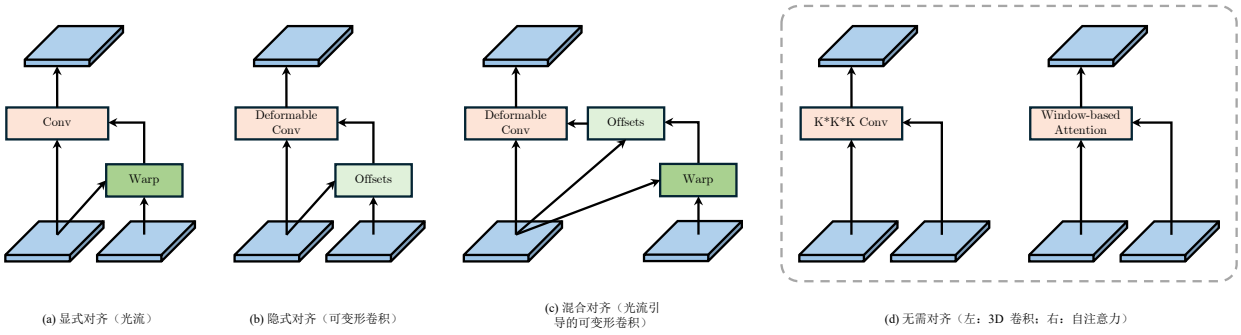

图 52　不同帧间对齐模式示意图

Fig. 52　Illustration of different inter-frame alignment

基于光流的显式对齐方法通过分析连续帧间的变化, 为每个像素计算运动矢量, 然后利用这些矢量信息通过双线性插值或空间变换网络等方式对相邻帧/特征进行变换, 以匹配支持帧像素在参考帧中的位置. BasicVSR[75]通过实验验证了光流估计不准确导致在图像层面的变形存在模糊、孔洞等问题, 降低了重建质量. 在特征层面的变形方式可以有效解决这一问题, BasicVSR 利用预训练的 SpyNet 估计相邻帧之间的运动矢量, 再通过双线性插值将由支持帧提取的特征对齐到参考帧. 然而, 依赖光流的显式对齐法只适用于短时间内运动幅度且光照变化较小的场景. 如果低分辨率的视频内容出现了较大的运动或明显的光照变化, 显式对齐法会显著降低 VSR 算法的性能.

隐式对齐法不依赖于明确的运动估计信息, 而是在特征层面自适应地建立帧间对应关系, 这类方法多利用具备了空间变换能力的可变形卷积 (Deformable convolution) [118–119] . 在传统的卷积操作中, 卷积核的形状是固定不变的. 而可变形卷积额外引入可学习的偏移量有助于在空间上自适应地调整卷积核, 以捕捉更复杂且更有价值的特征, 在一定程度上可以应对目标变形、尺度变化等造成的挑战, 如图 54 所示.

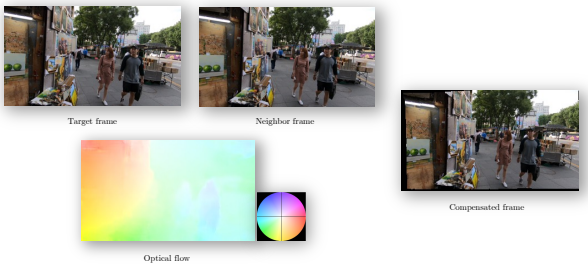

图 53　基于光流的显式运动对齐

Fig. 53　Explicit alignment based on optical flow

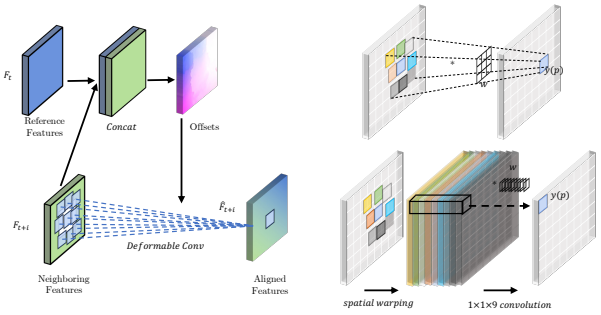

图 54 基于可变形卷积的对齐

Fig. 54 Deformable convolution-based alignment

在视频超分辨率的帧间对齐中, 传统的卷积运算局部感受野固定, 难以处理这些错综复杂的物体运动变化和形变. 而可变形卷积通过引入可学习的采样点偏移量, 使得卷积核可以适应物体在不同帧的位置偏移和外观形变, 获得更好的对齐效果, 提高了视频超分辨率重建的质量. 受 TDAN[51]的启发, EDVR[67]结合金字塔结构和可变形卷积, 由粗到细地将相邻帧的特征与参考帧的特征对齐, 以应对幅度大和复杂的运动变化. 与基于光流法建模相邻帧单一像素间的运动关系相比, 可变形卷积学习到的偏移量具有更强的探索性和更丰富的表现力, 可以更有效地避免不精确运动估计造成的伪影并为遮挡物体提供更多的细节信息. 然而, 基于可变形卷积的对齐方法在大模型的训练过程中极度不稳定, 会导致偏移量溢出. 当学习到的偏移量非常大时, 网络无法从相邻帧得到有效信息, 导致视频超分网络“退化”为一个单图超分网络.

混合对齐方法同时使用两种或两种以上的对齐方法, 结合了它们的优势应对不同类型的运动, 提高了帧间对齐的精度. Chan 等[120]揭示了基于可变形卷积的隐式对齐和基于光流的显式对齐之间的共同点, 认为可变形卷积的偏移多样性是为前者带来性能增益的关键, 提出了偏移保真损失函数 (Offset-fidelity loss). 该损失函数利用光流指导偏移量的学习过程, 解决了基于可变形卷积的对齐方法训练不稳定的问题. 在此基础上, BasicVSR++[83]设计了光流引导的可变形对齐, 一方面利用可变形卷积探索多个相关的像素以减少伪影, 另一方面采用光流作为初始偏移量, 减轻了可变形对齐模块的训练负担, 结构如图 55 所示.

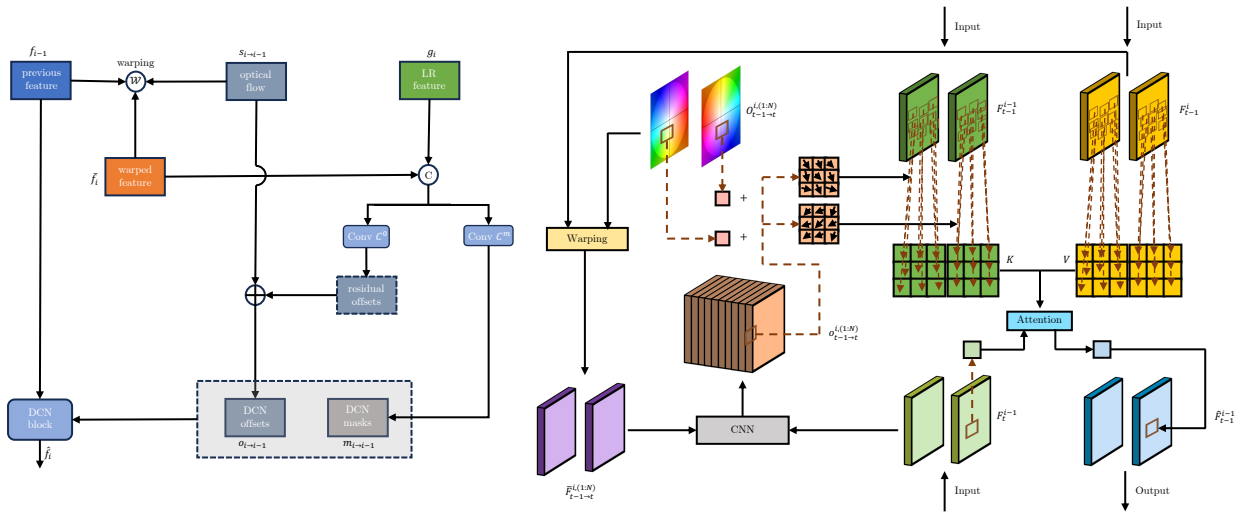

图 55 光流引导的可变形对齐和光流引导的可变形注意力

Fig. 55 Flow-guided deformable alignment and flow-guided deformable attention

在光流引导的可变形对齐中, 基于光流的预对齐是一次粗糙的对齐过程, 光流对齐的结果和光流一起用于学习可变形对齐中偏移量残差和幅度, 进而实现更加精细的对齐效果. 无独有偶, TTVSR[87]利用光流将视频帧的图像块沿时间维度串联, 形成多条预对齐的轨迹来描述视频中物体的运动, 只需利用注意力机制寻找位于同一轨迹中相似度最高的图像块作为当前帧的对齐帧, 能更加充分地利用较长视频序列的时序信息. RVRT[93]将视频序列分成多个片段, 将片段作为信息传播的基本单位达到了超分性能和计算复杂度之间的平衡. 在每个视频片段内, 并行提取所有帧的特征, 再通过自注意力机制隐式地融合和优化; 在不同的视频片段间, 利用光流引导的可变形注意力对齐片段间的信息, 融合时序信息. 考虑到不准确的光流估计和对齐过程中的插值操作破坏亚像素信息, PSRT[72]设计了联合光流的运动估计和不依赖插值的图像块对齐算法, 进一步提升了视频超分辨率重建算法的性能.

无需对齐法在特征融合或重建之前不进行帧间对齐操作, 而是通过网络建模视频中的时空相关性. 如图 56 所示, 3D 卷积、注意力计算等操作可以通过添加额外的维度或将多帧图像同时划分为视觉“令牌”的方式, 直接从原始未对齐的相邻帧中捕获亚像素信息重建当前帧. 以 3D 卷积、注意力计算为代表的“无需对齐法”可以处理具有各种光照和运动特性的视频, 具有较强的应用前景. 然而, 3D 卷积、注意力计算带来了巨大的计算开销, 限制了其广泛应用. DSMC[60]利用 3D 卷积直接处理连续而未对齐的多帧图像以应对视频内容大幅度的运动变化. 同时, 为了解决 3D 卷积计算复杂度高的问题, DSMC 对下采样的特征采用 3D 卷积操作提取特征, 并降低了特征的通道数量.

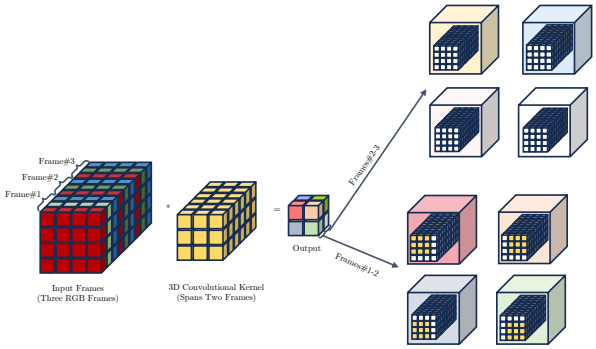

图 56　基于 3D 卷积的帧间对齐

Fig. 56　Inter-frame alignment based on 3D convolution

为了探究不同对齐方式对性能的影响, 本文以 VRT[88] 为骨干网络, 在合成数据集[121]上对比结合不同对齐方式所重建的视频质量, 如表 5 所示. 其中, 用于对齐的光流涵盖了合成数据集的真实光流和由 SpyNet 估计的光流信息. 实验结果表明, 在网络架构和特征提取模块相同的情况下, 混合对齐方式在光流先验的引导下能够获得更出色的重建质量, 但也不可避免地引入了更多的网络参数. 此外, 对比不同的插值方法和光流方法, 可以看出, 光流精度以及插值带来的误差对视频超分算法的性能影响较大. 例如, 光流引导的可变形卷积和光流引导的可变形注意力分别依赖于双线性和最近邻插值, 造成了重建的视频发生平滑亦或是失真现象, 导致其性能相对较弱. 光流引导的图像块对齐是针对不精确的光流估计以及插值引入的误差而设计的, 这意味着具有真实光流的合成数据集并非是该对齐方法的理想场景, 故其表现欠佳. 光流引导的隐式对齐借助注意力隐性地学习插值权重, 鲁棒性更强, 且在所有对齐过程中共享参数, 仅比光流引导的可变形卷积和可变形注意力增加了 0.01 M 模型参数.

表5　不同帧间对齐方式的性能和参数比较

Table 5　Performance and parameter comparisons of different inter-frame alignment

对齐方式	参数量 (M)	插值方法	光流	
			GT	SpyNet
显式对齐 (光流)	1.35	最近邻插值	31.84	31.78
		双线性插值	31.92	31.85
		双三次插值	31.93	31.89
混合对齐 (光流引导的可变形卷积)	1.60	双线性插值	32.08	31.98
混合对齐 (光流引导的可变形注意力)	1.56	双线性插值	32.03	31.94
混合对齐 (光流引导的图像块对齐)	1.35	最近邻插值	31.81	31.82‘
混合对齐 (光流引导的隐式对齐)	1.36	基于注意力的隐式插值	32.14	32.05

为了综合评测不同网络架构和帧间对齐模式给视频超分性能与效率带来的影响, 本文在同一平台 (GeForce RTX 3090) 下评估12种具有代表性的视频超分算法, 实验结果如表 6 所示. 其中, 网络的推理时间指的是将分辨率为 320 × 180 的低分辨率帧上采样至 1280 × 720 所需的时长, 即在 REDS 数据集上对连续 100 帧进行 4 倍上采样所需时间. 以对双三次插值下采样的 REDS 数据集的重建结果为例, 隐式对齐 (EDVR) 和无需对齐 (DUF) 相较于显式对齐 (TOFlow) 能够更有效地应对相邻帧的复杂运动和光照变化, 进而获得更佳的重建视频质量. 与此同时, 基于 3D 卷积和可变形卷积实现的非显式对齐方式显著增加了网络的参数量和推理时间. 相反, 基于单向传播 (TMP) 和基于双向传播 (BasicVSR) 的超分网络借助沿时间维度持续传递的隐状态, 凭借更少的网络参数对长程依赖关系进行建模, 在提升网络性能的同时降低了所需的推理时间. 同时, 随着网络结构由单向传播 (TMP) 向二阶网格结构 (BasicVSR++) 转变, 超分网络的传播分支逐渐增多, 获取的时序信息也更为充分, 相较于计算开销的增加性能提升更为显著.

然而, 在把网络中的卷积操作 (BasicVSR++) 替换成特征表示能力更强的自注意力操作 (PSRT) 之后, 网络的参数量和推理时间承受了巨大的负担, 但性能的提升也逐渐遭遇瓶颈. 值得关注的是, 混合对齐方法逐渐成为主流的帧间对齐模式, 这得益于其更为准确的对齐结果以及更稳定的训练. 此外, 与光流主导的显式对齐相比, 隐式对齐和无需对齐在面对复杂的运动模式和场景变换时具有更好的稳健性, 因此广泛应用于具有噪声等严重退化的真实场景的视频超分重建.

表6 GeForce RTX 3090 平台下 VSR 的性能和推理时间对比结果

Table 6 Performance and inference time comparisons of VSR algorithm on GeForce RTX 3090 platform

对比方法	参数量 (M)	推理时间 (ms)	对齐方式	双三次插值下采样			高斯模糊下采样		
				REDS (RGB 通道)	Vimeo-90K-T (Y 通道)	Vid4 (Y 通道)	Vimeo-90K-T (Y 通道)	Vid4 (Y 通道)	UDM10 (Y 通道)
Bicubic	-	< 1	-	26.23/0.7319	31.32/0.8684	23.78/0.6374	31.30/0.8687	21.80/0.5346	28.47/0.8253
TOFlow [12]	1.41	250	显式	27.96/0.7981	33.08/0.9054	25.89/0.7651	34.62/0.9212	25.85/0.7659	36.26/0.9438
DUF [56]	5.8	737.5	无需	28.63/0.8251	-/-	27.33/0.8319	36.87/0.9447	27.38/0.8329	38.48/0.9605
EDVR [67]	20.6	188.2	隐式	31.09/0.8800	37.61/0.9489	27.35/0.8264	37.81/0.9523	27.85/0.8503	39.89/0.9686
TMP [61]	3.1	31.5	隐式	30.67/0.8710	-/-	27.10/0.8167	37.33/0.9481	27.61/0.8428	-/-
BasicVSR [75]	6.3	45.4	显式	31.42/0.8909	37.18/0.9450	27.24/0.8251	37.53/0.9498	27.96/0.8553	39.96/0.9694
ICONVSR [75]	8.7	58.4	显式	31.67/0.8948	37.47/0.9476	27.39/0.8279	37.84/0.9524	28.04/0.8570	40.03/0.9694
TTVSR [87]	6.8	123.3	混合	32.12/0.9021	-/-	-/-	37.92/0.9526	28.40/0.8643	40.41/0.9712
VRT [88]	35.6	1679	混合	32.17/0.9002	38.20/0.9530	27.93/0.8425	38.72/0.9584	29.37/0.8792	41.04/0.9737
BasicVSR++ [83]	7.3	60.2	混合	32.39/0.9069	37.79/0.9500	27.79/0.8400	38.21/0.9550	29.04/0.8753	40.72/0.9722
PSRT [72]	13.4	1280.2	混合	32.72/0.9106	38.27/0.9536	28.07/0.8485	-/-	-/-	-/-
MIA-VSR [97]	16.5	1194.6	无需	32.78 0.9220	38.22/0.9532	28.20/0.8507	-/-	-/-	-/-

4 难点和未来研究

未来，基于深度学习的视频超分辨率重建研究方向主要包括：

1) 在线视频超分辨率重建

基于循环架构的方法可以利用整个视频序列的信息, 但在在线直播和视频会议等实际应用场景中, 超分网络仅能获得过去或有限数量的未来帧的信息, 限制了其重建视频的质量[122–123]. 同时, 参数庞大和计算复杂的网络模型难以适应在线视频要求实时超分的需求[124–125]. 此外, 在线视频复杂的视频内容和终端设备有限的计算资源也是在线 VSR 所面临的难点[126]. 因此, 作为视频超分辨率方法的重要特性, 如何有效利用不同帧中所含信息是在线视频超分重建任务亟待解决的问题之一[127–128].

2) 真实场景下的视频超分辨率重建

现有的视频超分模型主要在合成数据集上进行训练和测试的, 可以为低分辨率视频重建丢失的高频细节, 若将其直接应用于退化过程未知的视频, 会显著降低 VSR 性能. 这是因为合成数据集和真实场景下的退化过程存在较大的“域差异”, 在合成数据集上训练得到的模型无法完美泛化至真实场景[129]. 因此, 真实场景下的 VSR 面临的主要困境是缺乏大量且具有代表性的数据集. 通常, 真实场景的退化过程可能包括模糊[130]、噪声[131]、压缩伪影[132]、色彩失真等方式, 每种方式的不同形式也会使退化过程变得更复杂. 现有的帧间对齐方法可以应对插值和模糊等简单的退化过程[133–134], 对于非线性和复杂的运动、镜头畸变以及场景和光照变化等问题, 以光流为代表的对齐方法就显得力不从心了[135].

为了设计更稳健的视频超分算法, 仅依赖制作特定场景和退化过程的真实数据集是不够的, 还可以着眼于设计更合理的退化过程、引入无监督等技术. 例如, 在构建低分辨率视频时, 退化过程的建模在理论上应当与现实情形一致, 以减小研究与实践之间的差距[136]. 无监督的 VSR 算法[137–140] 也可以解决真实场景中成对数据获取难、成本高的问题. 此外, 现有的 VSR 数据集很少涵盖有场景显著变化的视频[141], 对存在场景变化的视频进行处理时, 就不得不将其分割为多个不存在场景变化的片段予以处理, 消耗巨大的计算资源, 在现实应用场景中, 也需要更多能够处理具有复杂场景变化视频的网络.

3) 任意上采样因子的视频超分辨率重建

受限于模型训练和评价数据集, 大多数基于深度学习的视频超分算法往往采用 4× 上采样因子, 无法适用于实际场景. 由于不同采集设备和显示设备之间的差异, 视频超分辨率重建的上采样因子并不一定是整数. 诸如 ×2、×3 或者 ×1.5 等缩放比例也极为常见. 同时, 随着大屏幕和超大屏幕显示设备的普及, 研究更大上采样因子的视频超分工作也具有巨大的研究价值和应用需求[142–143], 然而具有固定缩放比例的视频超分模型会严重制约其泛化能力与可移植性[144]. 因此, 如何在更大上采样因子的超分任务中抑制噪声、保持纹理细节也是下一步研究的难点之一.

4) 生成范式的视频超分辨率重建

由于扩散模型在图像和视频生成任务的优异表现, 基于生成模型的视频超分辨率重建算法也得到了广泛的关注[76,145–146]. 相较于基于监督学习的回归模型, 基于生成模型的 VSR 算法具有更强的内容

恢复和纹理生成能力, 这可以为用户带来更佳的视觉体验. 随着视频分辨率的提升和移动设备的普及, 视频超分技术也广泛应用于视频压缩和网络传输. 联合生成模型[147–148], 只需要传输下游任务或用户最关心的视觉内容, 余下部分可以交由生成模型恢复, 显著缓解了传输带宽的压力[149–150]. 因此, 如何利用生成模型提升视频超分重建性能也是下一步研究重点之一.

5) 结合语义感知的视频超分辨率重建

通常, 作为底层视觉任务的超分辨率重建算法被认为是目标检测、实例分割等高层任务的预处理技术. 经过超分辨率重建处理后, 原本图像的语义内涵没有改变, 而增加的细节将有助于提升下游任务的精度. 现有的 VSR 算法主要依赖从大量数据中拟合图像的退化过程, 忽视了对视频内容的理解, 无法准确恢复物体的结构和纹理信息. 然而, 真正有效的 VSR 算法应该是结合对图像内容的认知和先验知识[151–154], 以自上而下的方式修复低分辨率视频. 此外, 本文主要讨论了面向人类感知的视频超分辨率重建, 而结合语义感知的视频超分辨率重建, 也为面向机器视觉的视频超分辨率重建提供了细粒度的可操作性[155], 同样也是未来研究的难点之一.

6) 视频超分辩率重建网络的可解释性

在深度学习领域, 网络的可解释性[156–157]是一个非常重要且富有挑战性的研究内容. 对于视频超分任务而言, 如何深刻理解神经网络的工作机制及其内部结构对重建结果的影响, 将有助于获得更精确和清晰的视频超分结果[158]. 视频超分辨率重建网络可解释性研究包括但不限于：通道对网络性能和效率的影响、网络层特征与下游视觉任务特征间的区分和关联以及基于 Transformer 的网络中位置编码的作用等.

5 结束语

视频超分辨率重建是计算机视觉领域具有重要研究意义和应用价值的课题. 本文首先在深度学习技术的视角下对视频超分辨率重建任务进行了定义, 总结了常用的公共数据集. 然后, 根据不同的信息传播方式, 本文将基于深度学习的视频超分辨率重建算法分为基于并行架构的视频超分辨率重建算法和基于循环架构的视频超分辨率重建算法, 并梳理了相关算法的进展情况. 最后, 总结了目前 VSR 算法面临的挑战及下一步的研究思路.

References

1 Wan Z, Zhang B, Chen D, et al. Bringing old films back to life. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 17694−17703

2 Li G, Ji J, Qin M, et al. Towards high-quality and efficient video super-resolution via spatial-temporal data overfitting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 10259−10269

3 Zhu H, Wei Y, Liang X, et al. CTP: Towards vision-language continual pretraining via compatible momentum contrast and topology preservation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 22257−22267

4 Jiao S, Wei Y, Wang Y, et al. Learning mask-aware clip representations for zero-shot segmentation. In: Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS). New Orleans, USA: 2023. 35631−35653.

5 Liu C, Sun D. On Bayesian adaptive video super resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, **36**(2) : 346−360

6 Ma Z, Liao R, Tao X, et al. Handling motion blur in multi-frame super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 5224−5232

7 Wu Y, Li F, Bai H, et al. Bridging component learning with degradation modelling for blind image super-resolution. *IEEE Transactions on Multimedia*, DOI: 10.1109/TMM.2022.3216115

8 Zhang Shuai-Yong, Liu Mei-Qin, Yao Chao, Lin Chun-Yu, Zhao Yao. Hierarchical feature feedback network for depth super-resolution reconstruction. *Acta Automatica Sinica*, 2022, **48**(4) : 992−1003
(张帅勇, 刘美琴, 姚超, 林春雨, 赵耀. 分级特征反馈融合的深度图像超分辨率重建. 自动化学报, 2022, **48**(4) : 992−1003)

9 Charbonnier P, Blanc-Feraud L, Aubert G, et al. Two deterministic half-quadratic regularization algorithms for computed imaging. In: Proceedings of 1st International Conference on Image Processing (ICIP). Austin, USA: IEEE, 1994. 168−172

10 Lai W S, Huang J B, Ahuja N, et al. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **41**(11) : 2599−2613

11 Zha L, Yang Y, Lai Z, et al. A lightweight dense connected approach with attention on single image super-resolution. *Electronics*, 2021, **10**(11) : 1234

12 Xue T, Chen B, Wu J, et al. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 2019, **127**(8) : 1106−1125

13 Liu C, Sun D. A bayesian approach to adaptive video super resolution. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Colorado Springs, USA: IEEE, 2011. 209−216

14 Nah S, Baik S, Hong S, et al. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, USA: IEEE, 2019. 1996−2005

15 Protter M, Elad M, Takeda H, et al. Generalizing the nonlocal-means to super-resolution reconstruction. *IEEE Transactions on Image Processing*, 2008, **18**(1) : 36−51

16 O. Shahar, A. Faktor, and M. Irani, Space-time super-resolution from a single video. In: Proceedings of the IEEE Conference Computer Vision and Pattern Recognition (CVPR). Colorado Springs, USA: IEEE, 2011. 33533360

17 Li D, Wang Z. Video superresolution via motion compensation and deep residual learning. *IEEE Transactions on Computational Imaging*, 2017, **3**(4) : 749−762

18 Venice[Online], available: https://www.harmonicinc.com /free-4k-demo-footage/, May 1, 2017

19 Myanmar 60p, Harmonic Inc.[Online], available: http://www.harmonicinc.com/resources/videos/4k-video-clip-center, May 1, 2017

20 ITS, "Consumer digital video library" [Online], available: https://www.cdvl.org, March 20, 2024

21 Mercat A, Viitanen M, Vanne J. UVG dataset: 50/120fps 4K sequences for video codec analysis and development. In: Proceedings of the ACM Multimedia Systems Conference. Istanbul, Turkey: ACM, 2020. 297−302

22 Liu D, Wang Z, Fan Y, et al. Robust video super-resolution with learned temporal dynamics. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 2507−2515

23 Li D, Wang Z. Video superresolution via motion compensation and deep residual learning. *IEEE Transactions on Computational Imaging*, 2017, **3**(4) : 749−762

24 Wang Z, Yi P, Jiang K, et al. Multi-memory convolutional neural network for video super-resolution. *IEEE Transactions on Image Processing*, 2018, **28**(5) : 2530−2544

25 Yi P, Wang Z, Jiang K, et al. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019. 3106−3115

26 Yu J, Liu J, Bo L, et al. Memory-augmented non-local attention for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 17834−17843

27 Yang X, Xiang W, Zeng H, et al. Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 4781−4790

28 Chan K C K, Zhou S, Xu X, et al. Investigating trade-offs in real-world video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 5962−5971

29 Lee J, Lee M, Cho S, et al. Reference-based video super-resolution using multi-camera video triplets. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 17824−17833

30 Wang R, Liu X, Zhang Z, et al. Benchmark dataset and effective inter-frame alignment for real-world video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 1168−1177

31 Huang Y, Dong H, Pan J, et al. Boosting video super resolution with patch-based temporal redundancy optimization. In: Proceedings of International Conference on Artificial Neural Networks (ICANN). Heraklion, Greece: Springer, 2023. 362−375

32 Zhou S, Yang P, Wang J, et al. Upscale-A-Video: Temporal-consistent diffusion model for real-world video super-resolution. arXiv preprint arXiv:2312.06640, 2023.

33 Wang X, Xie L, Dong C, et al. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Montreal, Canada: IEEE, 2021. 1905−1914

34 Singh A, Singh J. Survey on single image based super-resolution−implementation challenges and solutions. *Multimedia Tools and Applications*, 2020, **79**(3-5) : 1641−1672

35 You Z, Li Z, Gu J, et al. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. arXiv preprint arXiv:2312.08962, 2023.

36 You Z, Gu J, Li Z, et al. Descriptive image quality assessment in the wild. arXiv preprint arXiv:2405.18842, 2024.

37 Xie L, Wang X, Zhang H, et al. VFHQ: A high-quality dataset and benchmark for video face super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 657−666

38 Zhou F, Sheng W, Lu Z, et al. A database and model for the visual quality assessment of super-resolution videos. *IEEE Transactions on Broadcasting*, 2024, **70**(2) : 516−532

39 Jin J, Zhang X, Fu X, et al. Just noticeable difference for deep machine vision. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021, **32**(6) : 3452−3461

40 Kappeler A, Yoo S, Dai Q, et al. Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging*, 2016, **2**(2) : 109−122

41 Lucas A, Lopez-Tapia S, Molina R, et al. Generative adversarial networks and perceptual losses for video super-resolution. *IEEE Transactions on Image Processing*, 2019, **28**(7) : 3312−3327

42 Caballero J, Ledig C, Aitken A, et al. Real-time video super-resolution with spatio-temporal networks and motion compensation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 4778−4787

43 Kim S Y, Lim J, Na T, et al. 3DSRNet: video super-resolution using 3D convolutional neural networks. arXiv preprint arXiv:1812.09079, 2018.

44 Li D, Liu Y, Wang Z. Video super-resolution using non-simultaneous fully recurrent convolutional network. *IEEE Transactions on Image Processing*, 2018, **28**(3) : 1342−1355

45 Haris M, Shakhnarovich G, Ukita N. Space-time-aware multi-resolution video enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 2859−2868

46 Bao W, Lai W S, Zhang X, et al. MEMC-Net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, **43**(3) : 933−948

47 Zhu X, Li Z, Lou J, et al. Video super-resolution based on a spatio-temporal matching network. *Pattern Recognition*, 2021, **110**: 107619

48 Wang L, Guo Y, Liu L, et al. Deep video super-resolution using HR optical flow estimation. *IEEE Transactions on Image Processing*, 2020, **29**: 4323−4336

49 Zhu X, Li Z, Zhang X Y, et al. Residual invertible spatio-temporal network for video super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Honolulu, USA: AAAI, 2019. 5981−5988

50 Bare B, Yan B, Ma C, et al. Real-time video super-resolution via motion convolution kernel estimation. *Neurocomputing*, 2019, **367**: 236−245

51 Tian Y, Zhang Y, Fu Y, et al. TDAN: Temporally-deformable alignment network for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 3360−3369

52 Ying X, Wang L, Wang Y, et al. Deformable 3D convolution for video super-resolution. *IEEE Signal Processing Letters*, 2020, **27**: 1500−1504

53 Yan B, Lin C, Tan W. Frame and feature-context video super-resolution. In: Proceedings of the 33th AAAI Conference on Artificial Intelligence (AAAI). Honolulu, USA: AAAI, 2019: 5597−5604

54 Liu S, Zheng C, Lu K, et al. Evsrnet: Efficient video super-resolution with neural architecture search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 2480−2485

55 Rota C, Buzzelli M, van de Weijer J. Enhancing perceptual quality in video super-resolution through temporally-consistent detail synthesis using diffusion models. arXiv preprint arXiv:2311.15908, 2023.

56 Jo Y, Oh S W, Kang J, et al. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE, 2018. 3224−3232

57 Yi P, Wang Z, Jiang K, et al. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019. 3106−3115

58 Sun W, Sun J, Zhu Y, et al. Video super-resolution via dense non-local spatial-temporal convolutional network. *Neurocomputing*, 2020, **403**: 1-12

59 Haris M, Shakhnarovich G, Ukita N. Recurrent back-projection network for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 3897−3906

60 Liu H, Zhao P, Ruan Z, et al. Large motion video super-resolution with dual subnet and multi-stage communicated upsampling. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Virtual Event: AAAI, 2021. 2127−2135

61 Zhang Z, Li R, Guo S, et al. TMP: Temporal motion propagation for online Video super-sesolution. arXiv preprint arXiv:2312.09909, 2023.

62 Li W, Tao X, Guo T, et al. Mucan: Multi-correspondence aggregation network for video super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020. 335−351

63 Song H, Xu W, Liu D, et al. Multi-stage feature fusion network for video super-resolution. *IEEE Transactions on Image Processing*, 2021, **30**: 2923−2934

64 Fuoli D, Danelljan M, Timofte R, et al. Fast online video super-resolution with deformable attention pyramid. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2023. 1735−1744

65 Kalarot R, Porikli F. Multiboot VSR: Multi-stage multi-reference bootstrapping for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, USA: IEEE, 2019. 2060−2069

66 Xia B, He J, Zhang Y, et al. Structured sparsity learning for efficient video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 22638−22647

67 Wang X, Chan K C K, Yu K, et al. EDVR: Video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, USA: IEEE, 2019. 1954−1963

68 Fuoli D, Gu S, Timofte R. Efficient video super-resolution through recurrent latent space propagation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul, Korea: IEEE, 2019. 3476−3485

69 Isobe T, Li S, Jia X, et al. Video super-resolution with temporal group attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 8008−8017

70 Jin S, Liu M, Yao C, et al. Kernel Dimension Matters: To activate available kernels for real-time video super-resolution. In: Proceedings of the ACM International Conference on Multimedia (ACM MM). Ottawa, Canada: ACM, 2023. 8617−8625

71 Cao J, Li Y, Zhang K, et al. Video super-resolution transformer. arXiv preprint arXiv:2106.06847, 2021.

72 Shi S, Gu J, Xie L, et al. Rethinking alignment in video super-resolution transformers. In: Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS). New Orleans, USA: 2022. 36081−36093

73 Tang Q, Zhao Y, Liu M, et al. SeeClear: Semantic distillation enhances pixel condensation for video super-resolution. arXiv preprint arXiv:2410.05799, 2024.

74 Huang C, Li J, Chu L, et al. Disentangle propagation and restoration for efficient video recovery. In: Proceedings of the ACM International Conference on Multimedia (ACM MM). Ottawa, Canada: ACM, 2023. 8336−8345

75 Chan K C K, Wang X, Yu K, et al. Basicvsr: The search for essential components in video super-resolution and beyond. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 4947−4956

76 Chen Z, Long F, Qiu Z, et al. Learning spatial adaptation and temporal coherence in diffusion models for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 9232−9241

77 Leng J, Wang J, Gao X, et al. Icnet: Joint alignment and reconstruction via iterative collaboration for video super-resolution. In: Proceedings of the ACM International Conference on Multimedia (ACM MM). Lisboa, Portugal: ACM, 2022. 6675−6684

78 Yi P, Wang Z, Jiang K, et al. A progressive fusion generative adversarial network for realistic and consistent video super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, **44**(5) : 2264−2280

79 Zhang F, Chen G, Wang H, et al. Multi-scale video super-resolution transformer with polynomial approximation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, **33**(9) : 4496−4506

80 Qiu Z, Yang H, Fu J, et al. Learning spatiotemporal frequency-transformer for compressed video super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV). Tel Aviv, Israel: Springer, 2022. 257−273

81 Jiang Y, Chan K C K, Wang X, et al. Reference-based image and video super-resolution via C^2-matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, **45**(7) : 8874−8887

82 Isobe T, Jia X, Tao X, et al. Look back and forth: Video super-resolution with explicit temporal difference modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 17411−17420

83 Chan K C K, Wang X, Yu K, et al. Basicvsr: The search for essential components in video super-resolution and beyond. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 4947−4956

84 Zhou K, Li W, Lu L, et al. Revisiting temporal alignment for video restoration. In: Proceedings/CVF of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 6053−6062

85 Tang Q, Zhao Y, Liu M, et al. Semantic lens: Instance-centric semantic alignment for video super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Vancouver, Canada: AAAI, 2024. 5154−5161

86 Liu M, Jin S, Yao C, et al. Temporal consistency learning of inter-frames for video super-resolution[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, **33**(4) : 1507−1520

87 Liu C, Yang H, Fu J, et al. Learning trajectory-aware transformer for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 5687−5696

88 Liang J, Cao J, Fan Y, et al. VRT: A video restoration transformer. *IEEE Transactions on Image Processing*, 2024, **33**: 2171−2182

89 Tang J, Lu C, Liu Z, et al. CTVSR: Collaborative spatial-temporal transformer for video super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, DOI: 10.1109/TCSVT.2023.3340439

90 Qiu Z, Yang H, Fu J, et al. Learning degradation-robust spatiotemporal frequency-transformer for video super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, **45**(12) : 14888−14904

91 Zhang C, Wang X, Xiong R, et al. Local-global dynamic filtering network for video super-resolution. *IEEE Transactions on Computational Imaging*, 2023, **9**: 963−976

92 Jiang L, Wang N, Dang Q, et al. PP-MSVSR: multi-stage video super-resolution. arXiv preprint arXiv:2112.02828, 2021.

93 Liang J, Fan Y, Xiang X, et al. Recurrent video restoration transformer with guided deformable attention. Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS). New Orleans, USA: 2022. 378−393

94 Dong S, Lu F, Wu Z, et al. DFVSR: directional frequency video super-resolution via asymmetric and enhancement alignment network. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). Macao, China: IJCAI, 2023. 681−689

95 Li F, Zhang L, Liu Z, et al. Multi-frequency representation enhancement with privilege information for video super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 12814−12825

96 Kai D, Lu J, Zhang Y, et al. EvTexture: Event-driven texture enhancement for video super-resolution. arXiv preprint arXiv:2406.13457, 2024.

97 Zhou X, Zhang L, Zhao X, et al. Video Super-Resolution Transformer with Masked Inter&Intra-Frame Attention. arXiv preprint arXiv:2401.06312, 2024.

98 Xu K, Yu Z, Wang X, et al. An implicit alignment for video super-resolution. arXiv preprint arXiv:2305.00163, 2023.

99 Huang Y, Wang W, Wang L. Video super-resolution via bidirectional recurrent convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **40**(4) : 1015−1028

100 Chu M, Xie Y, Mayer J, et al. Learning temporal coherence via self-supervision for GAN-based video generation. *ACM Transactions on Graphics*, 2020, **39**(4) : 75:1−75:13

101 Isobe T, Zhu F, Jia X, et al. Revisiting temporal modeling for video super-resolution. arXiv preprint arXiv:2008.05765, 2020.

102 Sajjadi M S M, Vemulapalli R, Brown M. Frame-recurrent video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE, 2018. 6626−6634

103 Isobe T, Jia X, Gu S, et al. Video super-resolution with recurrent structure-detail network. In: Proceedings of the European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020. 645−660

104 Baniya A A, Lee T K, Eklund P W, et al. Online video super-resolution using information replenishing unidirectional recurrent model. *Neurocomputing*, 2023, **546**: 126355

105 Lin J, Huang Y, Wang L. FDAN: Flow-guided deformable alignment network for video super-resolution. arXiv preprint arXiv:2105.05640, 2021.

106 Yi P, Wang Z, Jiang K, et al. Omniscient video super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 4429−4438

107 Li S, He F, Du B, et al. Fast spatio-temporal residual network for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 10522−10531

108 Yu J, Liu J, Bo L, et al. Memory-augmented non-local attention for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 17834−17843

109 Tao X, Gao H, Liao R, et al. Detail-revealing deep video super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 4472−4480

110 Yang X, He C, Ma J, et al. Motion-guided latent diffusion for temporally consistent real-world video super-resolution. arXiv preprint arXiv:2312.00853, 2023.

111 Liu H, Ruan Z, Zhao P, et al. Video super-resolution based on deep learning: a comprehensive survey. *Artificial Intelligence Review*, 2022, **55**(8) : 5981−6035

112 Tu Z, Li H, Xie W, et al. Optical flow for video super-resolution: A survey. *Artificial Intelligence Review*, 2022, **55**(8) : 6505−6546

113 Baniya A A, Lee G, Eklund P, et al. A methodical study of deep learning based video super-resolution. *Authorea Preprints*, DOI: 10.36227/techrxiv.23896986.v1

114 Jiang Jun-Jun, Cheng Hao, Li Zhen-Yu, Liu Xian-Ming, Wang Zhong-Yuan. Deep learning based video-related super-resolution technique: A survey. *Journal of Image and Graphics*, 2023, **28**(07) : 1927−1964
(江俊君, 程豪, 李震宇, 刘贤明, 王中元. 深度学习视频超分辨率技术概述. 中国图象图形学报, 2023, **28** (7) : 1927−1964)

115 Dong C, Loy C C, He K, et al. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, **38**(2) : 295−307

116 Drulea M, Nedevschi S. Total variation regularization of local-global optical flow. In: Proceedings of the International IEEE Conference on Intelligent Transportation Systems (ITSC). Washington, USA: IEEE, 2011. 318−323

117 Haris M, Shakhnarovich G, Ukita N. Deep back-projection networks for super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE, 2018. 1664−1673

118 Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 764−773

119 Zhu X, Hu H, Lin S, et al. Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 9308−9316

120 Chan K C K, Wang X, Yu K, et al. Understanding deformable alignment in video super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Virtual Event: AAAI, 2021: 973−981

121 Butler D J, Wulff J, Stanley G B, et al. A naturalistic open source movie for optical flow evaluation. In: Proceedings of European Conference on Computer Vision (ECCV). Florence, Italy: Springer, 2012. 611−625

122 Lian W, Lian W. Sliding window recurrent network for efficient video super-resolution. In: Proceedings of the European Conference on Computer Vision Workshops (ECCVW). Tel Aviv, Israel: Springer Nature Switzerland, 2022. 591−601

123 Xiao J, Jiang X, Zheng N, et al. Online video super-resolution with convolutional kernel bypass grafts. *IEEE Transactions on Multimedia*, 2023, **25**: 8972−8987

124 Li D, Shi X, Zhang Y, et al. A simple baseline for video restoration with grouped spatial-temporal shift. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 9822−9832

125 Geng Z, Liang L, Ding T, et al. Rstt: Real-time spatial temporal transformer for space-time video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 17441−17451

126 Lin L, Wang X, Qi Z, et al. Accelerating the training of video super-resolution models. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Washington, USA: AAAI, 2023. 1595−1603

127 Li H, Chen X, Dong J, et al. Collaborative feedback discriminative propagation for video super-resolution. arXiv preprint arXiv:2404.04745, 2024.

128 Hu M, Jiang K, Wang Z, et al. Cycmunet+: Cycle-projected mutual learning for spatial-temporal video super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, **45**(11) : 13376−13392

129 Xiao Y, Yuan Q, Jiang K, et al. Local-global temporal difference learning for satellite video super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, **34**(4) : 2789−2802

130 Hui Y, Liu Y, Liu Y, et al. VJT: A video transformer on joint tasks of deblurring, low-light enhancement and denoising. arXiv preprint arXiv:2401.14754, 2024.

131 Song Y, Wang M, Yang Z, et al. NegVSR: Augmenting negatives for generalized noise modeling in real-world video super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Vancouver, Canada: AAAI, 2024. 10705−10713

132 Wang Y, Isobe T, Jia X, et al. Compression-aware video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 2012−2021

133 Youk G, Oh J, Kim M. FMA-Net: Flow-guided dynamic filtering and iterative feature refinement with multi-attention for joint video super-resolution and deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 44−55

134 Zhang Y, Yao A. RealViformer: Investigating attention for real-world video super-resolution. arXiv preprint arXiv:2407.13987, 2024.

135 Xiang X, Tian Y, Zhang Y, et al. Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020: 3370−3379

136 Jeelani M, Cheema N, Illgner-Fehns K, et al. Expanding synthetic real-world degradations for blind video super resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Vancouver, Canada: IEEE, 2023. 1199−1208

137 Bai H, Pan J. Self-supervised deep blind video super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(7) : 4641−4653

138 Pan J, Bai H, Dong J, et al. Deep blind video super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR). Seattle, USA: IEEE, 2024. 4811−4820

139 Chen H, Li W, Gu J, et al. Low-res leads the way: Improving generalization for super-resolution by self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 25857−25867

140 Yuan J, Ma J, Wang B, et al. Content-decoupled contrastive learning-based implicit degradation modeling for blind image super-resolution. arXiv preprint arXiv:2408.05440, 2024.

141 Chen Y H, Chen S C, Lin Y Y, et al. MoTIF: Learning motion trajectories with local implicit neural functions for continuous space-time video super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 23131−23141

142 Huang C, Li J, Chu L, et al. Arbitrary-scale video super-resolution guided by dynamic context. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Vancouver, Canada: AAAI, 2024. 2294−2302

143 Li Z, Liu H, Shang F, et al. SAVSR: Arbitrary-scale video super-resolution via a learned scale-adaptive network. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Vancouver, Canada: AAAI, 2024. 3288−3296

144 Huang Z, Huang A, Hu X, et al. Scale-adaptive feature aggregation for efficient space-time video super-resolution. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2024. 4228−4239

145 Xu Y, Park T, Zhang R, et al. VideoGigaGAN: Towards detail-rich video super-resolution. arXiv preprint arXiv:2404.12388, 2024.

146 He Q, Wang S, Liu T, et al. Enhancing measurement precision for rotor vibration displacement via a progressive video super resolution network. *IEEE Transactions on Instrumentation and Measurement*, 2024, **73**: 1−13

147 Chang J, Zhao Z, Jia C, et al. Conceptual compression via deep structure and texture synthesis. *IEEE Transactions on Image Processing*, 2022, **31**: 2809−2823

148 Chang J, Zhang J, Li J, et al. Semantic-aware visual decomposition for image coding. *International Journal of Computer Vision*, 2023, **131**(9) : 2333−2355

149 Ren B, Li Y, Liang J, et al. Sharing key semantics in transformer makes efficient image restoration. arXiv preprint arXiv:2405.20008, 2024.

150 Wu R, Sun L, Ma Z, et al. One-step effective diffusion network for real-world image super-resolution. arXiv preprint arXiv:2406.08177, 2024.

151 Sun H, Li W, Liu J, et al. Coser: Bridging image and language for cognitive super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 25868−25878

152 Wu R, Yang T, Sun L, et al. Seesr: Towards semantics-aware real-world image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 25456−25467

153 Zhang Y, Zhang H, Chai X, et al. MRIR: Integrating multimodal insights for diffusion-based realistic image restoration. arXiv preprint arXiv:2407.03635, 2024.

154 Zhang Y, Zhang H, Chai X, et al. Diff-restorer: Unleashing visual prompts for diffusion-based universal image restoration. arXiv preprint arXiv:2407.03636, 2024.

155 Ouyang H, Wang Q, Xiao Y, et al. Codef: Content deformation fields for temporally consistent video processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024: 8089−8099

156 Hu J, Gu J, Yu S, et al. Interpreting low-level vision models with causal effect maps. arXiv preprint arXiv:2407.19789, 2024.

157 Gu J, Dong C. Interpreting super-resolution networks with local attribution maps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Virtual: IEEE, 2021. 9199−9208

158 Cao J, Liang J, Zhang K, et al. Towards interpretable video super-resolution via alternating optimization. In: Proceedings of the European Conference on Computer Vision (ECCV). Tel Aviv, Israel: Springer, 2022. 393−411

唐麒 北京交通大学信息科学研究所硕士研究生. 主要研究方向为图像与视频复原.
E-mail: qitang@bjtu.edu.cn
(**TANG Qi** Master student at Institute of Information Science, Beijing Jiaotong University. His research interest covers image and video restoration.)

赵耀 北京交通大学信息科学研究所教授. 主要研究方向为图像/视频压缩, 数字媒体内容安全, 媒体内容分析与理解, 人工智能.
E-mail: yzhao@bjtu.edu.cn
(**ZHAO Yao** Professor at Institute of Information Science, Beijing Jiaotong University. His research interest covers image/video compression, digital media content security, media content analysis and understanding, artificial intelligence.)

刘美琴 北京交通大学信息科学研究所教授. 主要研究方向为多媒体信息处理, 三维视频处理, 视频智能编码. 本文通信作者.
E-mail: mqliu@bjtu.edu.cn
(**LIU Mei-Qin** Professor at Institute of Information Science, Beijing Jiaotong University. Her research interest covers multimedia information processing, 3D video processing and video intelligent coding. Corresponding author of this paper.)

姚超 北京科技大学计算机与通信工程学院副教授. 主要研究方向为图像/视频压缩, 计算机视觉和人机交互.
E-mail: yaochao@ustb.edu.cn
(**YAO Chao** Associate Professor at School of Computer and Communication Engineering, University of Science and Technology Beijing. His research interest covers image/video compression, computer vision, and human – computer interaction.)