中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-15

论文引用格式:Wang Zhixiang, Zhang Yayuan, Shang Wei, Yang Liu, Zhu Pengfei, Ren Dongwei. Second-order alignment and laplacian pyramid priors for arbitrary-scale video super-resolution[J/OL]. Journal of Image and Graphics, XXXX:1-15. DOI: 10.11834/jig.250659.(王志翔, 张雅媛, 尚玮, 杨柳, 朱鹏飞, 任冬伟. 二阶对齐与频域先验引导的任意倍率视频超分辨[J/OL]. 中国图象图形学报, XXXX:1-15. DOI: 10.11834/jig.250659.)[DOI:10.11834/jig.250659]

二阶对齐与频域先验引导的任意倍率视频超分辨

王志翔[1],张雅媛[2],尚玮[3*],杨柳[1],朱鹏飞[1],任冬伟[1]

1. 天津大学智能与计算学部,天津　300354; 2. 天津大学精密仪器与光电子工程学院,天津　300072; 3. 新加坡管理大学计算机与信息系统学院,新加坡　178902

摘　要:**目的**　任意倍率视频超分辨(arbitrary-scale video super-resolution,AVSR)旨在根据指定倍率提升视频帧的空间分辨率。现有方法在细节恢复、时序一致性与计算效率之间仍存在权衡问题。**方法**　本文采用基于前瞻机制的循环神经网络作为整体框架,在兼顾性能与效率的基础上,融合多尺度频率先验、基于光流的传播单元、二阶可形变对齐单元和超上采样单元,以增强时空信息聚合及任意倍率重建能力。**结果**　在REDS数据集的多倍率测试中,本文方法相较代表性AVSR方法在PSNR上平均提升0.16 dB;在Vid4数据集的整数与非整数倍率测试中,仍表现出较好的跨数据集泛化能力。消融实验表明,二阶可形变对齐与多尺度频率先验能够有效提升复杂运动场景下的重建质量。**结论**　所提出的任意倍率视频超分辨方法能够兼顾重建精度、泛化能力与计算效率,为实际任意倍率超分应用提供了可行方案。本文代码已公开发布,相关资源可通过Science Data Bank获取:https://www.doi.org/10.57760/sciencedb.j00240.00181。

关键词:任意倍率视频超分辨;循环神经网络;二阶可形变对齐;频域先验;超上采样单元

Second-order alignment and laplacian pyramid priors for arbitrary-scale video super-resolution

Wang Zhixiang[1], Zhang Yayuan[2], Shang Wei[3*], Yang Liu[1], Zhu Pengfei[1], Ren Dongwei[1]

1. College of Intelligence and Computing, Tianjin University, Tianjin 300354 China; 2. School of Precision Instrument and Opto-Electronics Engineering, Tianjin University Tianjin 300354 China; 3. School of Computing and Information Systems, Singapore Management University, 178902 Singapore

Abstract: **Objective** Arbitrary-scale video super-resolution (AVSR) aims to reconstruct high-resolution (HR) videos from low-resolution (LR) inputs under continuous scaling factors, including non-integer and asymmetric magnifications. Compared with fixed-scale video super-resolution (VSR), AVSR must generalize across a continuum of scales while maintaining temporal coherence amid complex motions, non-rigid deformations, and occlusions. In practice, three key issues often drive performance degradation: (i) scale generalization, where details plausible at one magnification may appear over-smoothed or over-sharpened at another; (ii) alignment error accumulation, where minor misalignments from optical-flow warping compound during recurrent propagation, causing flickering, ghosting, and motion artifacts; and (iii) robustness to unseen degradations, as real videos often diverge from training degradation models, complicating high-frequency restora-

收稿日期:2025-12-30;**修回日期**:2026-04-13

*通信作者:尚玮 csweishang@gmail.com

基金项目:国家自然科学基金(62576241, 62172127, U22B2035)

Supported by:Project supported by the National Natural Science Foundation of China (62576241, 62172127, U22B2035)

tion and temporal stability. This work develops an AVSR approach that enhances spatial detail recovery, temporal consistency, and scale generalization while maintaining deployment-friendly efficiency. **Method** We propose SL-AVSR, an arbitrary-scale video super-resolution framework that integrates (1) an explicit multi-scale frequency prior derived from image Laplacian pyramids; (2) second-order composite-flow-guided propagation for temporal feature transfer; (3) second-order deformable alignment refinement for sub-pixel correction near motion boundaries and non-rigid regions; and (4) a scale-aware hyper-upsampling unit for efficient continuous scaling. SL-AVSR builds on a forward-looking recurrent architecture with a lightweight look-ahead mechanism. The current HR frame is reconstructed by fusing history-propagated features with a short window of future cues, avoiding the overhead of a full bidirectional pass. First, to ensure scale-consistent guidance for detail restoration, we construct a Laplacian pyramid on the LR input to extract band-limited components representing multi-scale frequency information. These components are fused via learnable weights, enabling the network to prioritize appropriate frequency bands for different magnifications and content types. Unlike resource-intensive perceptual feature networks, this explicit prior is lightweight, interpretable, and imposes direct constraints on frequency discrepancies across scales. Second, to enhance alignment robustness in recurrent temporal aggregation, SL-AVSR employs second-order composite flow for feature propagation. Instead of using one-step displacements from single neighboring frames, we compose neighboring flows into two-step composite displacements, providing more stable cues under large motions and partial occlusions. This composite-flow-guided warping transfers features temporally, mitigating drift and curbing misalignment error accumulation. Third, to resolve residual misalignments persisting after flow-based warping—particularly around motion boundaries, non-rigid deformations, and occlusions—we introduce a second-order deformable alignment refinement module. This module predicts residual sampling offsets and modulation masks conditioned on warped features and the current context, enabling adaptive local corrections around flow-estimated displacements. The refinement is applied in both history propagation and look-ahead aggregation pathways, improving temporal feature correspondence and reducing motion artifacts. Fourth, to enable efficient continuous and asymmetric scaling, SL-AVSR incorporates a scale-aware hyper-upsampling unit. A compact hyper-network generates scale-specific convolution kernels that can be precomputed or cached for common output resolutions. This approach balances (i) direct interpolation (fast but limited in fidelity for large scales and fine textures) and (ii) implicit neural representation (INR)-based pixel-wise rendering (flexible but computationally expensive). By conditioning convolutional kernels on the target scale, SL-AVSR preserves convolution-based efficiency alongside arbitrary-scale flexibility. **Result** Training occurs on standard VSR/AVSR benchmarks with continuous scale sampling, with evaluation under integer, non-integer, and asymmetric magnifications. Generalization is tested by applying models trained on one dataset directly to others without adaptation. Robustness is assessed under randomized synthetic degradations and real-world videos with unknown degradations. We report distortion metrics (PSNR, SSIM) and a perceptual metric (LPIPS) for fidelity and quality, alongside qualitative comparisons and time-space profile visualizations to evaluate temporal stability (e. g. , flickering and alignment artifacts). Across scaling factors (including non-integer and asymmetric) and diverse video content, SL-AVSR achieves the best or consistently competitive quantitative performance against representative AVSR and arbitrary-scale image super-resolution (AISR) baselines. The explicit Laplacian-pyramid frequency prior delivers stable gains in detail recovery and scale generalization, evidenced by higher PSNR/SSIM and lower LPIPS across most scales. Qualitatively, SL-AVSR reconstructs structured regions (e. g. , thin lines, repetitive patterns, man-made textures) more reliably and preserves stochastic textures with fewer over-smoothing artifacts, especially at large magnifications where frequency information is vulnerable. For temporal consistency, the second-order composite-flow-guided propagation and deformable alignment refinement reduce motion distortions like trailing edges, ghosting, and shimmering. Time-space profiles reveal smoother, more continuous traces in SL-AVSR compared to competitors′ blurred or jagged ones, indicating superior temporal aggregation. The look-ahead mechanism further boosts stability and perceptual quality by incorporating future context without a costly full-sequence backward pass. In cross-dataset tests, SL-AVSR sustains robust performance on unseen distributions, with gradual degradation as scaling increases. Under randomized and real-world degradations, it avoids severe artifact amplification, underscoring the resilience from explicit frequency guidance and second-order alignment. Efficiency analyses show SL-AVSR′s favorable quality-efficiency trade-off, outperforming INR-based methods due to its kernel-generating hyper-upsampling and lightweight prior. **Conclusion** We present SL-AVSR, an

arbitrary-scale video super-resolution framework that unifies an explicit Laplacian-pyramid multi-scale frequency prior with second-order composite-flow-guided propagation and second-order deformable alignment refinement in a forward-looking recurrent architecture. The proposed design enhances spatial detail restoration and scale generalization while improving temporal consistency by mitigating alignment error accumulation under challenging motion patterns. The hyper-upsampling unit supports continuous scaling with practical efficiency, avoiding the high computational cost of pixel-wise implicit rendering. Extensive evaluations across datasets, scaling factors, and degradation conditions demonstrate SL-AVSR's strong balance of fidelity, perceptual quality, temporal coherence, and computational efficiency, positioning it as a practical solution for real-world arbitrary-scale video super-resolution. The code is publicly available through Science Data Bank: https://www. doi. org/10. 57760/sciencedb. j00240. 00181.

Key words: arbitrary-scale video super-resolution; recurrent neural network; second-order deformable alignment; frequency prior; hyper-upsampling unit

0 引　言

受限于成像设备、传输带宽和存储成本等因素，实际采集到的视频往往分辨率较低，难以满足超高清显示、视频监控和智能交通等场景对细节分辨能力的需求。视频超分辨(video super-resolution, VSR)通过利用相邻帧的时序冗余，从低分辨率(low resolution, LR)视频重建高分辨率(high resolution, HR)序列，是缓解上述矛盾的重要途径。Kappeler等人(2016)较早将卷积神经网络引入VSR，将多帧配准与深度重建统一到端到端框架中，验证了深度学习在视频重建中的潜力。Caballero等人(2017)提出利用运动补偿与时空卷积的实时视频超分网络，在保证推理速度的同时提升了重建质量。Wang等人(2019)提出的EDVR网络采用金字塔级可形变对齐和时空注意机制，在复杂运动和模糊场景下取得了显著性能提升。Chan等人(2021, 2022)提出BasicVSR与BasicVSR++，系统分析了传播、对齐和聚合等关键组件，在多个基准数据集上取得领先性能，并成为后续视频复原研究的重要基线。靳雨桐等(2022)提出轻量级注意力约束对齐网络，通过可变形对齐与动态融合实现高质量视频超分重建。江俊君等(2023)对基于深度学习的视频超分辨率技术进行了系统综述，为理解现有VSR方法的研究脉络与关键问题提供了参考。这些方法多针对固定整数超分倍率设计，通常需要针对不同超分倍率分别训练模型，当实际应用中存在连续变焦或任意倍率放大的需求时，难以在重建质量与部署成本之间取得理想平衡。

为提升超分辨灵活性，任意倍率超分辨(arbitrary-scale super-resolution)逐渐成为研究热点。在图像场景中，Meta-SR引用元上采样模块根据尺度因子与坐标动态预测卷积核，使单一网络即可处理任意实数超分倍率；ArbSR引用尺度自适应模块作为“插件”嵌入现有固定尺度网络，实现从多尺度网络向任意倍率网络的知识迁移。Chen等(2021)提出局部隐式图像函数(local implicit image function, LIIF)，将图像表示为坐标到像素值的连续映射，利用局部特征与坐标输入的多层感知机实现任意倍率下的像素预测。与单幅图像任意倍率超分相比，任意倍率视频超分辨(arbitrary-scale video super-resolution, AVSR)不仅需要统一建模连续空间尺度，还需在时间维度上对复杂运动进行对齐与聚合。一方面，Chen等(2022)在VideoINR中将视频编码为时空隐式神经表示，使网络能够在连续的时空坐标上查询任意分辨率和帧率的输出，从而统一解决时空超分问题。另一方面，MoTIF(Chen等，2023)在局部隐式神经函数框架下显式学习像素级前向运动轨迹，将运动建模与隐式表示结合，用于连续时空视频超分。然而，这类方法通常依赖复杂的隐式求值过程和有限的时间邻域，推理效率和长时序建模能力在高分辨率场景中仍然受限。近期方法同样致力于解决这些问题，如SAVSR(Li等, 2024)采用滑动窗口双向RNN提升时序聚合，但受限于窗口大小导致效率低下；ST-AVSR(Shang等, 2024)通过光流引导循环单元和多尺度结构-纹理先验增强细节恢复与尺度泛化；BF-STVSR(Kim等, 2025)引入双向流动时空变压器建模长时序依赖，进一步

缓解复杂运动下的误差。

基于上述分析，本文提出一种结合二阶对齐与拉普拉斯金字塔频率先验的任意倍率视频超分辨方

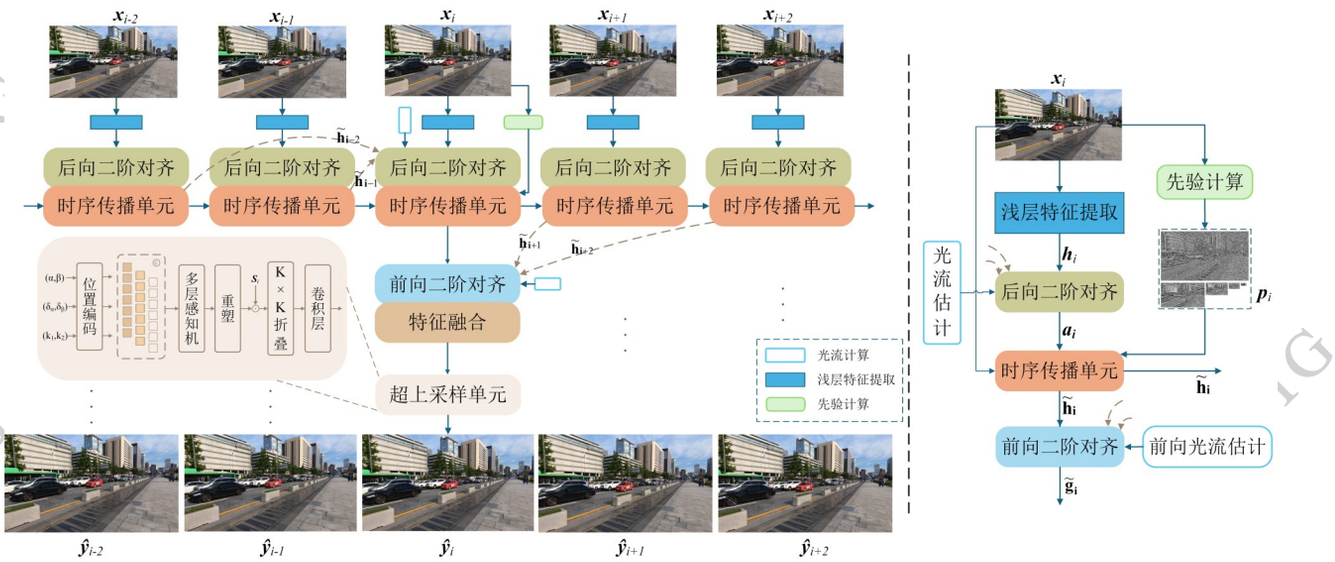

图 1　所提方法的总体框架。

Fig. 1　Overall framework of the proposed method.

法SL-AVSR，针对复杂运动下对齐误差累积和高频细节恢复不足的问题，在保持整体结构简洁高效的前提下进一步增强细节恢复与时序一致性。本文的主要贡献如下：

1）提出了一种基于前瞻机制的循环神经网络框架，用于实现任意倍率的视频超分辨率任务。该框架融合了基于图像拉普拉斯金字塔的自适应多尺度先验、光流引导的传播单元、二阶可形变对齐单元以及超上采样单元，有效实现了双向循环神经网络重建性能与单向循环神经网络计算效率之间的平衡。

2）提出了基于图像拉普拉斯金字塔生成的自适应多尺度先验模块，从不同频带提取结构-纹理信息，并自适应注入重建过程，有效提升大倍率超分条件下的空间细节恢复与尺度泛化能力。

3）将二阶可形变对齐单元引入基于前瞻机制的循环神经网络框架中，通过相邻帧与跨帧复合光流共同引导的可形变对齐，实现更精细的相邻帧与跨帧空间对齐，并有效聚合过去与未来的多帧信息。

4）在多个公开视频超分数据集及复杂退化设置下开展充分实验验证。结果表明，本方法可以兼顾重建精度、时序一致性与计算效率，且在不同超分倍率与退化强度下均表现出稳定的性能提升。

1　本文方法

给定一个低分辨率（LR）视频序列 $\boldsymbol{x}=\{\boldsymbol{x}_i\}_{i=0}^{T}$，$\boldsymbol{x}_i\in\mathbf{R}^{3\times H\times W}$，式中 $\boldsymbol{x}_i$ 为第 i 帧，H 和 W 分别表示帧的高与宽。本文提出的SL-AVSR的目标是重建一个高分辨率（HR）视频序列 $\hat{\boldsymbol{y}}=\{\hat{\boldsymbol{y}}_i\}_{i=0}^{T}$，$\hat{\boldsymbol{y}}_i\in\mathbf{R}^{3\times(\alpha H)\times(\beta W)}$，式中 $\alpha,\beta\geqslant 1$ 是用户指定的两个超分倍率。我们的方法SL-AVSR由四个模块构成：1）多尺度频率先验：提供与内容相关的像素级先验线索以引导复原；2）跨帧光流引导传播单元：利用相邻光流与跨帧复合光流对多帧隐藏状态进行对齐与聚合，建模长时程时空依赖并缓解对齐误差累积；3）二阶可形变对齐单元：在初步光流对齐的基础上，进一步预测光流残差与遮挡掩码，实现由粗到细的亚像素级精确对齐，该模块可同时用于聚合历史信息与邻近未来信息；4）任意倍率重建与上采样单元：生成可预计算的超分倍率相关的卷积核以实现任意倍率视频超分。整体系统框架图及部分数据流如图1所示，左侧为视频超分辨的整体流程，右侧为当前帧超分辨的详细过程。为简化图示，先验计算等重复操作仅在代表性帧位置示意，其余帧采用相同计算路径。

融合二阶可形变对齐与多尺度频率先验，是面向任意倍率视频超分核心挑战的协同设计。二者分别对应AVSR的两类关键瓶颈：时序误差累积与空间细节丢失。任意倍率超分需要在连续尺度下重建视频，使模型在时间维度需适应更复杂的运动形态（如大位移、非刚性形变），在空间维度则需从有限观

测中恢复跨尺度高频细节。为此，二阶可形变对齐通过复合光流引导与残差偏移细化，抑制长时序传播中由遮挡与运动估计误差引起的对齐漂移；拉普拉斯金字塔先验通过显式分解多频率信息，为不同放大倍率提供尺度自适应的结构—纹理约束。两者分别从时域稳定性与空域保真度两个互补维度建模，共同提升任意倍率重建的可靠性。

具体而言，二阶对齐输出的时序聚合特征具有更少的运动伪影与错位误差，为多频带先验融合提供了更准确的对齐基础，使先验中的高频成分能够注入到真实边缘位置，而非落入错位残差区域；同时，拉普拉斯先验提供的显式频率信息（尤其是高频纹理线索）可强化纹理边界等细节区域的对齐敏感性，引导残差偏移预测更精细，降低过度平滑导致的对齐退化。上述双向促进机制使模型在复杂运动场景下同时保持时序一致性与纹理清晰度，从而获得超越单一模块叠加的性能增益。

1.1 多尺度频率先验

对任意倍率视频超分而言，准确刻画图像在多尺度上的结构与纹理至关重要。计算机视觉与图像处理中的尺度空间理论为此提供了优雅的理论框架。由于图像的不同频带能够表征其结构和纹理，本文采用由拉普拉斯金字塔分解得到的多尺度频率先验，来替代我们此前工作中使用的VGG特征。该方法将图像分解为多个不同频带的层级；得到的金字塔包含若干层，每一层对应一个尺度并捕获特定的频率信息。不同频带能显式提供细节与纹理在不同尺度下的空间分布信息。通过用拉普拉斯金字塔分解得到的频域信息替代深度学习特征，我们在不依赖预训练网络的情况下获得了相当的性能，从而减少参数量与推理时间。

具体而言，我们将不同频带上采样到与输入分辨率一致，并对每个频带施加可学习权重进行逐频带加权融合，使网络能够针对不同视频自适应调整各频带的重要性。随后，将融合后的频率图与当前帧$\boldsymbol{x}_i$在通道维拼接，作为多尺度频率先验，记为$\boldsymbol{p}_i$。该先验通过以下方式引入模型：除生成高分辨率输出$\hat{\boldsymbol{y}}$的最终残差连接外，将模型内部所有源自初始输入$\boldsymbol{x}$的特征，均替换为对应的多尺度先验，以实现先验信息的全面注入。在图1框架中，我们从当前视频帧$\boldsymbol{x}_i$提取$\boldsymbol{p}_i$，并在时序传播单元和特征融合部分进行注入。

1.2 跨帧光流引导传播单元

给定LR视频序列$\boldsymbol{x}=\{\boldsymbol{x}_i\}_{i=0}^{T}$，光流引导循环单元用于计算隐藏状态序列$\{\boldsymbol{h}_i\}_{i=0}^{T}$，以建模历史帧的长时程时空依赖。设flow(·)为预训练且在训练过程中参数冻结的光流估计网络（实现中采用PWC-Net），warp(·)为采用双线性核的特征扭曲算子。本文以$\boldsymbol{f}_{a\to b}$表示从第a帧到第b帧的光流。首先，对相邻帧估计光流

$$\boldsymbol{f}_{i\to i-1}=\mathrm{flow}\left(\boldsymbol{x}_i,\boldsymbol{x}_{i-1}\right),i\in\{1,2,\cdots,T\},\#(1)$$

式中$\boldsymbol{f}_{i\to i-1}\in\mathbf{R}^{2\times H\times W}$。为将上一时刻隐藏状态对齐到当前帧，我们使用该光流对$\tilde{\boldsymbol{h}}_{i-1}$进行扭曲：

$$\boldsymbol{h}_{i-1\to i}=\mathrm{warp}\left(\tilde{\boldsymbol{h}}_{i-1},\boldsymbol{f}_{i\to i-1}\right),\#(2)$$

并将初始隐藏状态置零。仅依赖相邻帧对齐时，模型在大位移或遮挡场景下容易出现对齐误差累积。为此，本文进一步显式构造跨帧的复合光流$\boldsymbol{f}_{i\to i-2}$来刻画残余运动，需要强调的是$\boldsymbol{f}_{i\to i-2}$由相邻光流复合得到，而非直接估计：

$$\boldsymbol{f}_{i\to i-2}=\boldsymbol{f}_{i\to i-1}+\mathrm{warp}\left(\boldsymbol{f}_{i-1\to i-2},\boldsymbol{f}_{i\to i-1}\right),i\geqslant 2.\#(3)$$

由于双线性扭曲难以覆盖复杂非刚性运动以及遮挡边界处的局部错位，我们进一步引入二阶可形变对齐模块$A(\cdot)$对粗对齐结果进行细化。该模块以隐藏状态$\tilde{\boldsymbol{h}}_{i-1}$，$\tilde{\boldsymbol{h}}_{i-2}$，与当前帧$\boldsymbol{x}_i$输入网络的浅层卷积网络$\phi(\cdot)$得到的初始隐藏状态$\boldsymbol{h}_i=\phi(\boldsymbol{x}_i)$为条件（图1右侧浅层特征提取模块），并联合相邻光流与跨帧复合光流$\boldsymbol{f}_{i\to i-1}$，$\boldsymbol{f}_{i\to i-2}$预测可形变卷积的残差偏移与调制掩码，从而在光流轨迹附近自适应修正采样位置，得到历史聚合特征（具体方法见1.3节）

$$\boldsymbol{a}_i=A\left(\tilde{\boldsymbol{h}}_{i-1},\tilde{\boldsymbol{h}}_{i-2},\boldsymbol{h}_i;\boldsymbol{f}_{i\to i-1},\boldsymbol{f}_{i\to i-2}\right).\#(4)$$

此处光流扭曲与二阶可形变对齐对应图1中的二阶对齐模块。

记第i帧对应的多尺度频率先验为$\boldsymbol{p}_i$，则循环单元在第i个时间步的输入为

$$\boldsymbol{z}_i=\left[\boldsymbol{h}_i,\boldsymbol{p}_i,\boldsymbol{a}_i\right],\#(5)$$

式中[·]表示通道维拼接。随后将$\boldsymbol{z}_i$输入由N_1个残差块构成的卷积网络$\mathrm{F}_{\mathrm{rnn}}(\cdot)$更新隐藏状态：

$$\tilde{\boldsymbol{h}}_i=\mathrm{F}_{\mathrm{rnn}}\left(\boldsymbol{z}_i\right),i=0,1,\cdots,T.\#(6)$$

即图1中的时序传播单元部分。在序列起始阶段，当$i<2$无法形成跨帧复合光流时，可令跨帧分

支退化为相邻分支(如复用$\boldsymbol{f}_{i\to i-1}$及其对应的对齐特征),从而保持传播过程的连续性。

为兼顾双向循环神经网络对未来帧信息利用的优势与单向循环神经网络的推理效率,本研究引入前瞻机制作为时序建模的核心框架。该机制是一种轻量级的时序信息聚合策略,定义为:在单向前向的循环推理过程中,对当前帧的局部未来邻域进行特征提取与对齐聚合,在不构建完整反向传播分支的前提下,为当前帧重建补充局部未来时序线索,实现“前向推理 + 局部未来感知”的折中设计。具体操作为:在当前帧i之后取长度为$L=2$的滑动窗口,并采用与历史分支相同的二阶对齐策略:由未来窗口内的相邻光流复合得到指向当前帧的跨帧光流,并将未来隐藏状态对齐到当前帧坐标系;随后复用二阶可形变对齐模块A(·)得到前视聚合特征

$$\tilde{\boldsymbol{g}}_i = \mathrm{A}\left(\tilde{\boldsymbol{h}}_{i+1},\tilde{\boldsymbol{h}}_{i+2},\tilde{\boldsymbol{h}}_i;\boldsymbol{f}_{i\to i+1}\boldsymbol{f}_{i\to i+2}\right),\#(7)$$

式中$\tilde{\boldsymbol{h}}_i$为已经聚合历史信息的隐藏状态,$\boldsymbol{f}_{i\to i+t}$为由相邻光流复合得到的跨帧光流。$\tilde{\boldsymbol{g}}_i$与当前帧和先验经$N_2$个残差块特征融合得到$\boldsymbol{g}_i$,$\boldsymbol{g}_i$作为后续上采样分支的补充特征,用于增强细节恢复。

1.3 二阶可形变对齐模块

本小节对1.2节中的二阶可形变对齐模块A(·)进行细节阐述。在得到了相邻与跨帧的光流后,即可利用它们将多帧隐藏状态对齐到当前帧并进行聚合。为便于叙述,以后向对齐为例,模块首先利用双线性扭曲得到粗对齐条件特征

$$\boldsymbol{h}_{i-1\to i} = \mathrm{warp}\left(\tilde{\boldsymbol{h}}_{i-1},\boldsymbol{f}_{i\to i-1}\right),$$
$$\boldsymbol{h}_{i-2\to i} = \mathrm{warp}\left(\tilde{\boldsymbol{h}}_{i-2}\boldsymbol{f}_{i\to i-2}\right),\#(8)$$

并将其与初始隐藏状态$\boldsymbol{h}_i$拼接,连同光流一起作为偏移预测网络的输入:

$$\boldsymbol{e} = \left[\boldsymbol{h}_{i-1\to i},\boldsymbol{h}_i,\boldsymbol{h}_{i-2\to i}\boldsymbol{f}_{i\to i-1}\boldsymbol{f}_{i\to i-2}\right].\#(9)$$

随后,通过轻量卷积网络$\Psi(\cdot)$预测可形变卷积的残差偏移与调制掩码。残差偏移使用tanh(·)进行幅值约束,并乘以最大残差系数η以稳定训练:

$$\Delta\boldsymbol{o} = \eta\cdot\tanh\left(\Psi_{\mathrm{off}}(\boldsymbol{e})\right),\boldsymbol{m} = \mathrm{sigmoid}\left(\Psi_{\mathrm{mask}}(\boldsymbol{e})\right).\#(10)$$

最终偏移由基于光流的基础位移与网络预测的残余位移叠加得到:将$\boldsymbol{f}_{i\to i-1}$、$\boldsymbol{f}_{i\to i-2}$按可形变卷积的通道组织方式广播到偏移张量维度后,与$\Delta\boldsymbol{o}$相加得到$\boldsymbol{o}$。在实现中,我们将偏移预测网络最后一层卷积初始化为零,使得训练初期网络只依赖预训练的光流引导,初期网络训练过程更稳定。

对齐与融合阶段采用调制可形变卷积算子D(·)。在通道维拼接的张量$\boldsymbol{c} = [\tilde{\boldsymbol{h}}_{i-1},\tilde{\boldsymbol{h}}_{i-2}]$作为被采样输入,使用偏移$\boldsymbol{o}$与掩码$\boldsymbol{m}$在光流轨迹邻域内自适应选取采样位置并完成聚合,输出二阶对齐特征

$$\boldsymbol{a}_i = \mathrm{D}(\boldsymbol{c};\boldsymbol{o},\boldsymbol{m}).\#(11)$$

对齐过程以光流作为基础位移,并通过残差偏移在其邻域内进行修正,结合掩码抑制不可靠的采样位置。由此得到的二阶对齐特征能在保持运动一致性的同时更好地适配复杂局部形变与遮挡变化,为后续重建提供更稳定的跨帧信息支持。

1.4 超上采样单元

受神经克里金上采样器(neural kriging upsampler)(Wang 等,2023)的启发,本文设计的超上采样单元由两条分支组成:超分辨特征准备和超分辨卷积核生成,如图1左下所示。对于超分辨特征准备分支,我们将前视二阶对齐单元输出的特征$\boldsymbol{g}_i$与当前隐藏状态$\boldsymbol{h}_i$在通道维上拼接,并通过残差网络得到用于重建的SR特征。随后,将每个空间位置周围$K\times K$邻域内的C维SR特征展开为$C\times K^2$个通道,即为图像处理中img2col(·)操作在张量上的推广。最后,对展开后的特征采用双线性插值上采样至目标分辨率,得到目标图$\boldsymbol{s}_i$。

对于超分辨卷积核生成分支,我们训练一个超网络,即带有周期激活函数的多层感知机(multilayer perceptron,MLP)(Chen 等,2023),用于预测上采样卷积核$\boldsymbol{w}$。已有研究表明,周期激活能够有效缓解MLP的频谱偏置问题,在函数拟合能力上优于ReLU非线性(Sitzmann 等,2020)。为了使生成的卷积核具有尺度感知而与内容无关的特性,我们精心设计了MLP的输入,包括:1)超分倍率(α,β);2)LR帧与HR帧之间的相对坐标$(\delta_\alpha,\delta_\beta)$;3)卷积核$\boldsymbol{w}$的空间索引$(k_1,k_2)$。其中,前两项已在其他连续表示方法中得到应用(Chen 等,2021;Lee 等,2022)。为增强与尺度相关输入的判别性,我们在送入MLP之前对上述输入施加正弦型位置编码。需要指出的是,对于不同的目标分辨率,所对应的上采样卷积核$\boldsymbol{w}$可以预先计算并缓存,从而加速推理过程。

在获得卷积核$\boldsymbol{w}$之后,我们先对$\boldsymbol{w}$与特征$\boldsymbol{s}_i$进行Hadamard逐元素乘法,然后通过折叠操作,即展开操作的逆过程,将其还原回空间特征图。接着,采

用1×1卷积在通道维上进行信息融合，并使用3×3卷积进行通道调整，两者之间插入LeakyReLU激活函数。最后，将该3×3卷积层输出与双线性上采样后的LR帧逐像素相加，得到最终的HR帧$\hat{\boldsymbol{y}}_i$。

2 实　验

本节首先介绍实验设置，然后将所提出的SL-AVSR与当前代表性的任意倍率图像超分辨（AISR）和任意倍率视频超分辨（AVSR）方法进行对比，最后通过一系列消融实验验证SL-AVSR各关键设计的有效性。

图2 不同方法在REDS数据集上的视觉结果对比图

2.1 实验设置

2.1.1 数据集

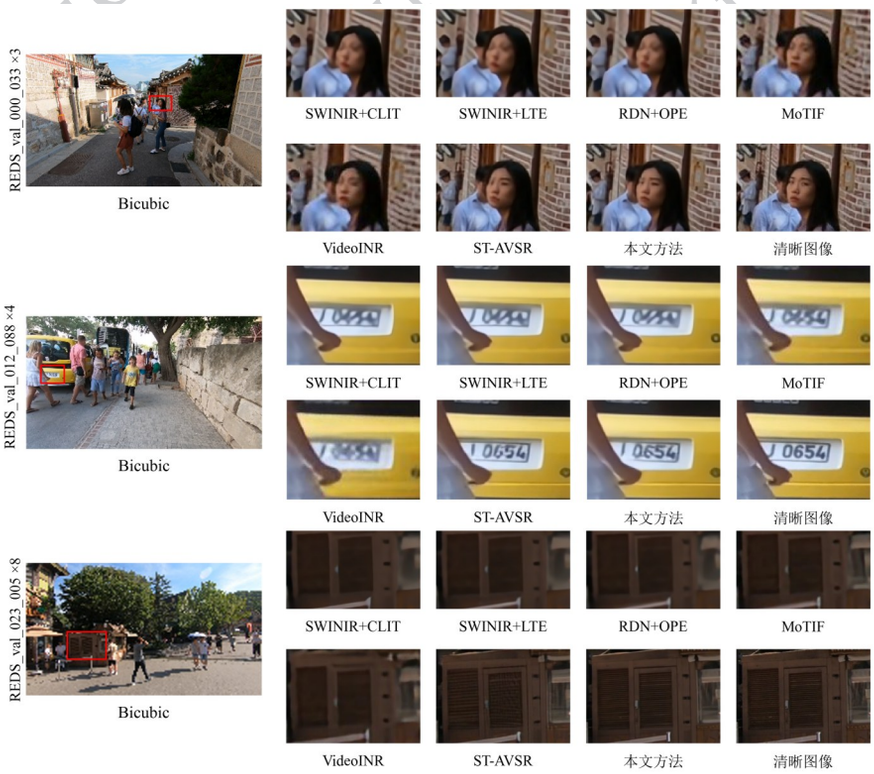

图3 Vid4数据集上不同超分倍率下PSNR与LPIPS指标的变化。

SL-AVSR在REDS数据集（Nah 等，2019）上训练，该数据集包含240个由GoPro捕获的分辨率为720×1280的视频。每个视频由100个HR帧组成。按照参考文献（Chen 等，2023）、（Chen 等，2023）、（Chen 等，2022）实验中的设置，我们使用双三次退化模型生成LR帧，并从均匀分布$U[1,4]$中随机采样超分倍率(α,β)。我们在包含30个视频的REDS验证集和包含4个视频的Vid4数据集（Liu 和 Sun，2013）上测试SL-AVSR。

为了评估我们的方法对未见退化模型的泛化能力，我们将视频随机退化管道（Chan 等，2022）应用于GoPro（Nah 等，2017）的测试集，结合噪声和视频压缩来合成未见的退化用于验证我们的方法的泛化能力。

2.1.2 数据预处理

为支持LR/HR分辨率均可变化的mini-batch训练，我们将EQSR（Wang 等，2023）中针对AISR的数据预处理策略扩展到AVSR任务。具体而言，对于尺寸为$\alpha P\times\beta P\times T$的HR视频块，我们首先将

其缩放至$P\times P\times T$，作为输入的LR视频块。随后，从同一HR视频块中裁剪出一组尺寸为$P\times P\times T$的真实标签视频块。对于同一LR输入对应的不同HR标签块，我们记录其各自的相对坐标$(\delta_\alpha,\delta_\beta)$，在超上采样单元中利用这些坐标实现对相

表 1　REDS 验证集上的定量比较(PSNR ↑ / SSIM ↑ / LPIPS ↓)。最佳结果以粗体突出显示。

Table 1　Quantitative comparison with state-of-the-art methods on the REDS validation set (PSNR ↑ / SSIM ↑ / LPIPS ↓), The best results are highlighted in bold.

方法		×2	×3	×4	×6	×8
Bicubic		31.51/0.911/0.165	26.82/0.788/0.377	24.92/0.713/0.484	22.89/0.622/0.631	21.69/0.574/0.699
EDVR		36.03/0.961/0.072	32.59/0.904/0.108	30.24/0.853/0.202	27.02/0.733/0.349	25.38/0.678/0.411
ArbSR		34.48/0.942/0.096	30.51/0.862/0.200	28.38/0.799/0.295	26.32/0.710/0.428	25.08/0.641/0.492
EQSR		34.71/0.943/0.082	30.71/0.867/0.194	28.75/0.804/0.283	26.53/0.718/0.391	25.23/0.645/0.459
RDN	+LTE	34.63/0.942/0.093	30.64/0.865/0.204	28.65/0.801/0.289	26.46/0.714/0.410	25.15/0.660/0.488
	+CLIT	34.63/0.942/0.092	30.63/0.865/0.204	28.63/0.801/0.290	26.43/0.714/0.400	25.14/0.661/0.467
	+OPE	34.05/0.939/0.082	30.52/0.864/0.199	28.63/0.800/0.293	26.37/0.711/0.421	25.04/0.655/0.504
	+GaussianSR	34.25/0.940/0.091	30.56/0.866/0.201	28.64/0.800/0.291	26.40/0.712/0.419	25.08/0.657/0.501
	+ContinuousSR	---/---/---	30.65/0.866/0.198	28.67/0.801/0.289	26.49/0.715/0.402	25.14/0.662/0.470
SwinIR	+LTE	34.73/0.943/0.091	30.73/0.866/0.200	28.75/0.804/0.284	26.56/0.718/0.403	25.24/0.669/0.480
	+CLIT	34.63/0.942/0.093	30.64/0.865/0.205	28.64/0.802/0.291	26.45/0.715/0.400	25.15/0.662/0.466
	+OPE	33.39/0.935/0.081	29.40/0.820/0.217	28.49/0.785/0.292	26.30/0.698/0.398	25.01/0.648/0.487
	+GaussianSR	34.31/0.941/0.089	30.60/0.867/0.199	28.69/0.802/0.290	26.42/0.713/0.416	25.08/0.659/0.498
	+ContinuousSR	---/---/---	30.75/0.868/0.197	28.68/0.805/0.287	26.58/0.720/0.401	25.26/0.670/0.467
VideoINR		31.59/0.900/0.144	30.04/0.852/0.197	28.13/0.791/0.263	25.27/0.687/0.374	23.46/0.619/0.470
MoTIF		31.03/0.898/0.100	30.44/0.862/0.186	28.77/0.807/0.260	25.63/0.698/0.369	25.12/0.664/0.467
BF-STVSR		32.06/0.908/0.092	31.38/0.877/0.146	29.29/0.837/0.200	25.98/0.718/0.321	25.42/0.670/0.459
SAVSR		35.66/0.955/0.046	32.19/0.918/0.100	30.61/0.872/0.138	27.03/0.791/0.250	25.59/0.716/0.312
ST-AVSR		36.91/0.969/0.041	33.41/0.937/0.066	31.03/0.897/0.114	27.89/0.812/0.222	26.04/0.746/0.298
本文方法		**37.09/0.970/0.041**	**33.68/0.940/0.065**	**31.27/0.900/0.114**	**28.01/0.814/0.220**	**26.15/0.747/0.293**

应区域的局部超分,从而使得网络能够支持在单批次中同时训练不同超分倍率的样本。数据增强方面,我们采用随机旋转(90°、180°或270°)以及随机水平翻转和垂直翻转。

2.1.3　具体设置

SL-AVSR 采用端到端方式优化,共训练 300K 次迭代。优化器选用 Adam(Kingma 和 Ba,2014),初始学习率为 2×10^{-4},并通过余弦退火策略(Loshchilov 和 Hutter,2017)逐渐衰减至 1×10^{-6}。除非另有说明,输入 LR 图像块尺寸设为 $P = 80$,序列长度为 $T = 15$,滑动窗口大小为 $L = 2$,残差块个数设为 $N_1 = N_2 = 15$,展开邻域大小为 $K = 3$,SR 特征通道数为 $C = 64$。超上采样单元中 MLP 的隐藏层维度依次设为 16、16、16 和 64。作为光流估计器的 PWC-Net(Ranjan 和 Black,2017)保持参数冻结。损失函数采用 Charbonnier 损失(Lai 等,2017):

$$l(\hat{\boldsymbol{y}},\boldsymbol{y}) = \frac{1}{(T+1)|\mathcal{Z}|}\sum_{t=0}^{T}\sum_{z\in\mathcal{Z}}\sqrt{(\hat{\boldsymbol{y}}_t(z)-\hat{\boldsymbol{y}}_t(z))^2+\epsilon}, \tag{12}$$

式中,$\hat{\boldsymbol{y}}$ 与 $\boldsymbol{y}$ 分别表示重建视频与真实视频,$\mathcal{Z}$ 为像素位置集合,ϵ 为平滑项。

2.2　与现有方法的比较

我们将 SL-AVSR 与近期具有代表性的 AISR 与 AVSR 方法进行对比。针对 AISR,我们选取三类代表方法:1) 基于可学习自适应滤波/核函数的任意倍率重建方法,包括 ArbSR(Wang 等,2021)与 EQSR(Wang 等,2023);2) 基于隐式神经表示的连续尺度查询方法,包括 LTE(Lee 和 Jin,2022)、CLIT(Chen 等,2023)与 OPE(Song 等,2023);3)基于高斯溅射的连续表示方法,包括 GaussianSR(Hu 等,2025)与 ContinuousSR(Peng 等,2025)。针对 AVSR,我们与 VideoINR(Chen 等,2022)、MoTIF(Chen 等,2023)、SAVSR(Li 等,2024)、ST-AVSR(Shang 等,2024)以及 BF-STVSR(Kim 等,2025)进行比较。为保证对比的公平性,我们在 REDS 数据集上对所有竞争方法进行统一微调;随后在 Vid4(Liu 和 Sun,2013)的整数与非整数倍率下评估跨数据集的泛化能力,并进一步在 GoPro 上采用随机退化策略(Chan 等,2022)生成的合成退化序列以及真实场景数据,评估模型对未知退化的鲁棒性。

2.2.1　在 REDS 数据集上的比较

受益于二阶对齐与多尺度频率先验,SL-AVSR

表 2　Vid4数据集上的定量比较(PSNR ↑ / SSIM ↑ / LPIPS ↓)。分数线上下分别表示水平、垂直方向的超分倍率，最佳结果以粗体突出显示。

Table 2　Quantitative comparison with state-of-the-art methods on the Vid4 set (PSNR ↑ / SSIM ↑ / LPIPS ↓). The scale factors above and below the score line indicate super-resolution in the horizontal and vertical directions, respectively. The best results are highlighted in bold.

方法		$\times\frac{2.5}{3.5}$	$\times\frac{4}{4}$	$\times\frac{5.1}{6}$	$\times\frac{6.4}{9}$
Bicubic		23.00/0.728/0.396	20.96/0.617/0.498	19.34/0.508/0.659	18.15/0.430/0.732
ArbSR		25.86/0.815/0.224	24.01/0.721/0.313	22.03/0.602/0.413	20.34/0.515/0.498
EQSR		26.24/0.826/0.210	24.16/0.730/0.300	22.54/0.620/0.399	**20.81**/0.528/0.472
RDN	+LTE	25.98/0.818/0.226	24.03/0.722/0.312	22.41/0.614/0.409	20.60/0.522/0.480
	+CLIT	25.83/0.815/0.223	23.94/0.721/0.312	22.38/0.613/0.411	20.57/0.520/0.491
	+OPE	25.77/0.818/0.217	23.98/0.719/0.317	22.35/0.610/0.416	20.55/0.528/0.495
	+GaussianSR	25.81/0.817/0.222	23.99/0.720/0.313	22.40/0.613/0.410	20.56/0.520/0.484
	+ContinuousSR	25.94/0.820/0.216	24.08/0.725/0.310	22.44/0.615/0.408	20.69/0.525/0.473
SwinIR	+LTE	26.43/0.826/0.217	24.09/0.727/0.305	22.50/0.620/0.403	20.70/0.524/0.475
	+CLIT	25.89/0.818/0.224	24.00/0.724/0.314	22.42/0.617/0.406	20.69/0.522/0.479
	+OPE	25.55/0.801/0.221	23.93/0.711/0.320	22.40/0.612/0.412	20.65/0.520/0.492
	+GaussianSR	25.92/0.820/0.220	24.01/0.722/0.311	22.44/0.615/0.407	20.66/0.523/0.480
	+ContinuousSR	26.54/0.830/0.210	24.16/0.729/0.301	22.52/0.622/0.401	20.80/0.539/0.469
VideoINR		23.02/0.715/0.203	24.34/0.741/0.249	22.02/0.601/0.397	20.43/0.511/0.453
MoTIF		23.55/0.734/0.209	24.52/0.746/0.261	22.11/0.604/0.390	20.48/0.518/0.450
BF-STVSR		24.12/0.745/0.166	24.90/0.784/0.222	22.23/0.620/0.388	20.59/0.537/0.447
SAVSR		27.82/0.875/0.088	25.97/0.835/0.154	22.40/0.679/0.348	20.73/0.588/0.393
ST-AVSR		29.09/0.913/**0.069**	26.16/0.852/0.127	23.02/0.735/0.253	20.64/**0.609/0.357**
本文方法		**29.38/0.917/0.069**	**26.23/0.859/0.123**	**23.41/0.742/0.245**	20.66/**0.609/0.357**

在表1所列的各项指标(PSNR/SSIM/LPIPS)及全部超分倍率上均取得最优结果。就AISR方法而言ContinuousSR在已报告的倍率上表现出较强的整体性能;但在×2设置下易出现条纹伪影,故其×2的结果未在表中给出。进一步比较可见,代表性的AVSR方法(如SAVSR、ST-AVSR)整体优于AISR方法,表明显式的时序建模对视频恢复至关重要。SAVSR采用窗口内双向RNN进行时序聚合,递归式推理限制了并行度,因而在性能与效率之间存在折中。ST-AVSR通过长序列建模并引入结构与纹理先验获得更好的重建质量;在此基础上,SL-AVSR进一步提升了细节重建的准确性:相较ST-AVSR,在×2/×3/×4/×6/×8上的PSNR分别提升0.10/0.23/0.21/0.14/0.12 dB(平均约0.16 dB)。

如图2所示,相比AISR方法,AVSR方法通常可产生更少伪影且更一致的时序细节。其中SL-AVSR在不同尺度下能更稳定地恢复细节并抑制失真,在复杂运动区域,如×3倍率样例中的人物面部,由于转头等非刚性运动,对比方法的重建结果容易出现边界模糊或五官畸变。相比之下,本文方法对人脸轮廓及五官结构进行了更贴近真实的还原,表现出更强的运动鲁棒性。在结构与纹理重建方面,×4倍率样例中的车牌区域显示,SL-AVSR恢复的字符更为清晰锐利,显著减少了其他方法中常见的笔画缺失或冗余伪影。特别是在×8极端倍率下,面对门上的扇叶窗这类重复性纹理,本文方法有效保持了结构的完整性,且未产生竖向条纹等失真,展现了优越的高倍率泛化能力。

2.2.2　在Vid4数据集上的泛化性

在REDS上训练的模型可直接迁移至Vid4,用于检验跨数据集的泛化能力。表2给出了在整数和非整数倍率下的定量测试结果,表明SL-AVSR在不同缩放设置下整体表现最佳:在多数尺度上,PSNR、SSIM与LPIPS均达到最优或次优水平;少数超分倍率下个别对比方法在PSNR上略占优势,但整体差距较小。图3进一步展示了不同超分倍率下PSNR与LPIPS的变化趋势。我们观察到,MoTIF在非整数及非对称尺度下的性能稳定性较弱,PSNR曲线呈现

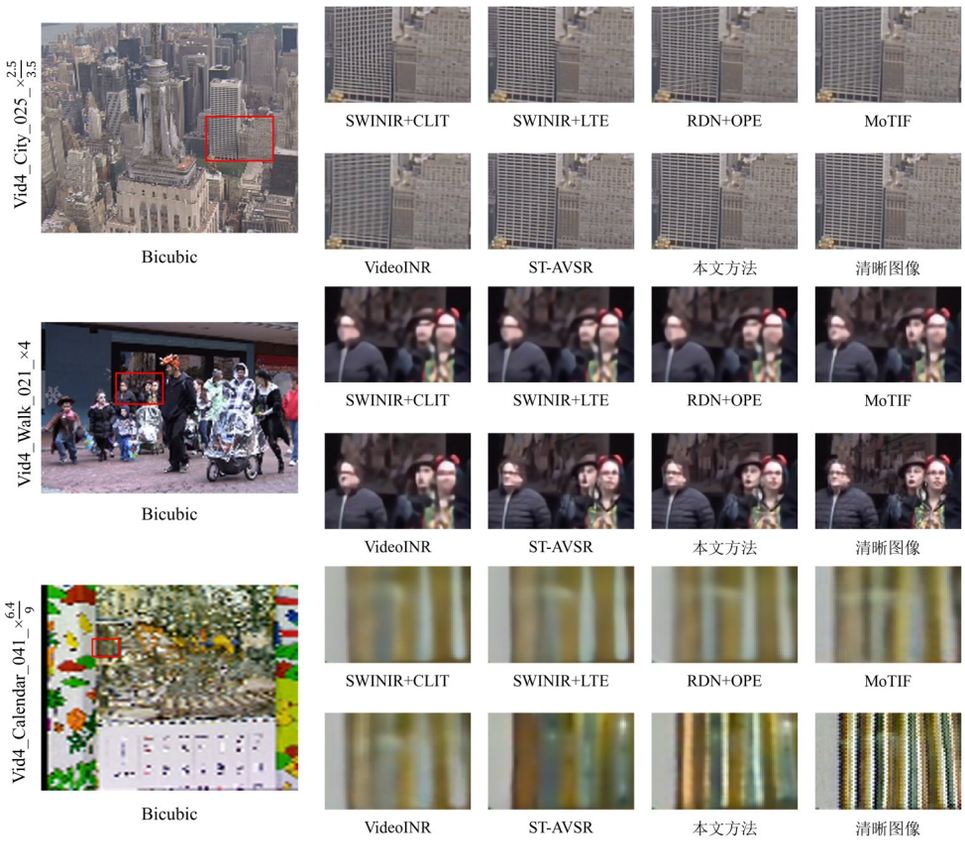

图 4　Vid4数据集上不同AVSR方法的视觉效果对比。

Fig. 4　Visual comparison of different AVSR methods on Vid4.

更明显的波动；我们推测该现象与超分结果与参考帧之间的亚像素级对齐误差有关，从而导致基于像素级误差的PSNR对错位更敏感。相比之下，LPIPS依赖深层特征度量，对小幅空间错位相对不敏感，因此对应波动不如PSNR显著。对于SL-AVSR，其在整数、非整数及非对称超分倍率下的性能随尺度增大呈现更平滑的下降趋势，体现出较好的尺度泛化能力。效率方面，各方法在GoPro的×4设置下的平均推理时间见表3（该实验在NVIDIA RTX Ada 5880上进行），SL-AVSR明显快于基于隐式神经表示的对比方法。

图4展示了Vid4数据集上的主观重建效果对比。可以看出，SL-AVSR在非结构化纹理与规则结构纹理区域均表现出更少的重影现象，并能更好地保持边缘连续性与纹理一致性。在×2.5/×3.5倍率下，建筑物的矩形窗口区域在重建时易产生各向异性伪影。本文方法有效抑制了此类伪影的产生，使窗口结构更加规整，显著提升了视频重建的视觉流畅度与时序一致性。在×4倍率样例中，对于面部等精细结构，本文方法重建的人脸五官清晰可辨，未发生明显形变；相比其他方法，五官区域之间没有出现粘连或模糊现象，较好地恢复了人脸的整体轮廓。在×6.4/×9高超分倍率下，面对密集条纹图案，SL-AVSR重建的条纹界限更加清晰，尤其白色条纹得以准确恢复，未与其他颜色的条纹发生混叠或融合，展现出在极端尺度下对高频细节的稳定保持能力。

图5进一步从时间维度评估各方法的时序一致性。具体而言，沿图中红色虚线所示的列抽取像素值并沿时间轴堆叠，即可得到时空剖面图，用于直观反映视频序列的时序稳定性。观察可见，多数对比方法生成的剖面呈现出更明显的模糊或锯齿状纹路，反映出较强的时间闪烁与不稳定。相比之下，SL-AVSR的剖面更为连续平滑，说明其时序聚合更为稳健，具有更强的时序一致性保持能力。

2.2.3　对未见退化模型的泛化

一个实用的AVSR方法必须在各种可能未见的退化下有效。为了评估这一点，我们通过结合更复杂的视频退化（Chan 等，2022）来生成测试视频序列，例如在双三次下采样之前添加噪声和视频压缩，这些在训练数据中不存在。我们将上述管道应用于

GoPro的测试集，以创建一个具有未见退化的测试集。以×4超分辨率为例，我们对所有方法在参数、计算成本和运行时间方面进行了全面比较，使用NVIDIA RTX Ada5880 GPU，结果如表3所示。我们的方法展示了比现有方法优越的泛化能力，不仅在处理未见退化模型时达到了性能的最优，而且在保持参数量和复杂度轻量化的同时，推理时间上显著优于竞争方法。

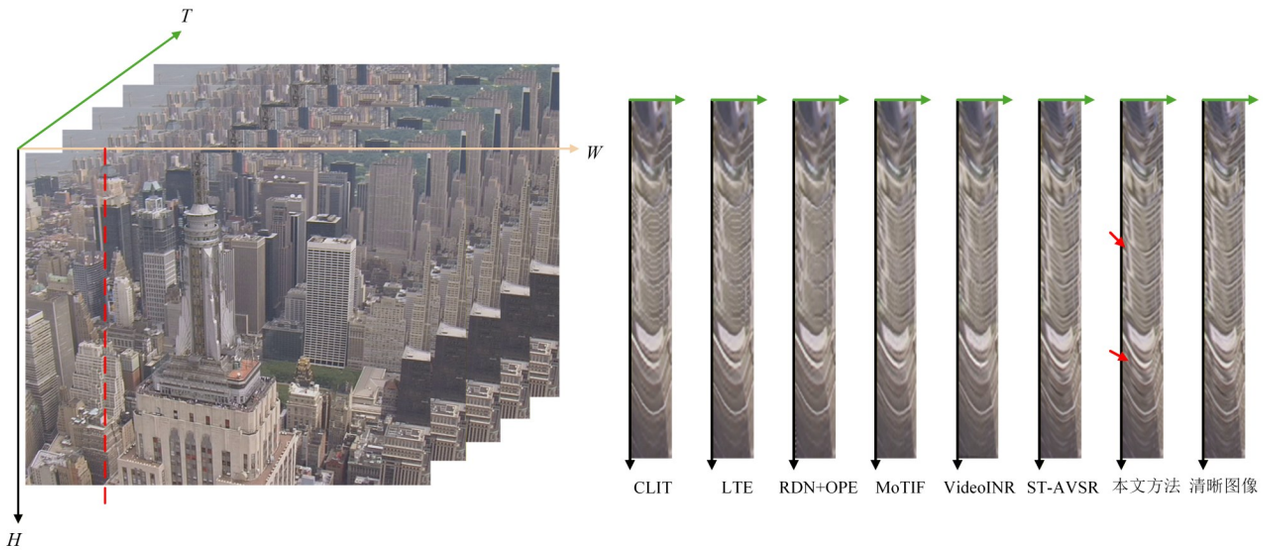

表3　GoPro数据集上×4超分辨率在未知退化下的泛化性能比较，以及参数量、计算复杂度和推理时间对比。

Table 3　Comparison of the generalization on GoPro for ×4 SR under unseen degradations, along with an efficiency comparison in terms of parameters, complexity, and inference time.

方法		PSNR ↑ / SSIM ↑ / LPIPS ↓	参数量(M)	复杂度(GFLOPs)	推理时间 (s)
Bicubic		23.63/0.711/0.416	--	--	--
ArbSR		27.43/0.798/0.239	16.6	887.3	0.651
EQSR		28.00/0.815/0.228	11.6	1743.2	0.921
RDN	+LTE	28.02/0.805/0.233	22.5	2011.3	0.519
	+CLIT	28.02/0.805/0.238	37.7	7341.9	1.655
	+OPE	27.90/0.798/0.242	22.1	1003.7	0.266
	+GaussianSR	27.97/0.801/0.240	23.2	1576.4	0.712
	+ContinuousSR	28.04/0.805/0.236	26.0	1980.1	0.319
SwinIR	+LTE	28.09/0.806/0.231	12.1	1692.8	0.729
	+CLIT	28.10/0.806/0.237	27.3	7022.3	1.928
	+OPE	28.02/0.802/0.240	11.7	684.0	0.438
	+GaussianSR	28.06/0.804/0.236	12.8	1257.9	0.923
	+ContinuousSR	28.09/0.807/0.233	15.6	1661.6	0.502
VideoINR		27.89/0.802/0.221	11.3	1676.5	0.676
MoTIF		28.02/0.810/0.219	12.6	2826.2	1.132
BF-STVSR		28.14/0.812/0.213	13.5	1876.4	1.003
SAVSR		29.67/0.849/0.193	11.5	1148.0	0.817
ST-AVSR		29.70/0.852/0.195	27.9	**296.8**	0.101
本文方法		**29.81/0.854/0.191**	**3.28**	354.0	**0.087**

图5不同方法的时序一致性对比。

AVSR方法通常比AISR方法具有更多参数和计算复杂度，因为它们利用相邻帧来增强当前帧的重建。VideoINR、MoTIF和BF-STVSR由于在训练和推

理过程中采用连续表示方法而计算开销较大。SAVSR使用滑动窗口进行传播，这导致了较高的计

上采样		先验			二阶对齐		超分倍率					推理时间(s)
①	②	①	②	③	①	②	×2	×3	×4	×6	×8	
✓				✓	✓	✓	36.30/0.965/0.047	32.75/0.935/0.083	30.58/0.895/0.135	27.09/0.806/0.254	25.38/0.739/0.321	0.053
	✓	✓			✓	✓	36.88/0.967/0.042	33.49/0.939/0.068	31.10/0.897/0.118	27.87/0.809/0.226	25.98/0.742/0.297	0.082
	✓		✓		✓	✓	36.98/0.968/0.041	33.62/0.940/0.066	31.23/0.900/0.114	28.01/0.813/0.221	26.13/0.747/0.294	0.104
	✓			✓	✓		36.67/0.967/0.044	32.86/0.927/0.081	30.23/0.880/0.147	27.36/0.786/0.243	25.58/0.718/0.314	0.071
	✓			✓		✓	36.76/0.967/0.043	32.97/0.933/0.079	30.56/0.883/0.136	27.55/0.797/0.239	25.79/0.730/0.307	0.076
	✓			✓	✓	✓	37.09/0.970/0.041	33.68/0.940/0.065	31.27/0.900/0.114	28.01/0.814/0.220	26.15/0.747/0.293	0.087

表5　不同滑动窗口长度L在REDS数据集上的消融实验结果(PSNR↑ / SSIM↑ / LPIPS↓)。

Table 5　Ablation results of different sliding window lengths L on the REDS dataset (PSNR↑ / SSIM↑ / LPIPS↓).

窗口长度	超分倍率				
	×2	×3	×4	×6	×8
L=0	36.41/0.965/0.046	32.88/0.929/0.082	30.61/0.886/0.137	27.42/0.795/0.251	25.63/0.729/0.321
L=1	36.76/0.968/0.044	33.24/0.934/0.074	30.89/0.892/0.128	27.67/0.803/0.234	25.79/0.734/0.309
L=2	37.09/0.970/0.041	33.68/0.940/0.065	31.27/0.900/**0.114**	28.01/**0.814/0.220**	**26.15/0.747/0.293**
L=3	**37.13/0.971/0.040**	**33.73/0.941/0.062**	**31.31/0.901**/0.116	**28.02**/0.812/0.224	26.09/0.742/0.301

算成本和推理时间。相比之下，我们的SL-AVSR在性能和效率之间取得了良好的平衡。

表4 SL-AVSR在REDS数据集上的消融分析(PSNR↑ / SSIM↑ / LPIPS↓)。不同变体的详细说明见正文。

Table 4 Ablation analysis of SL-AVSR on REDS (PSNR↑ / SSIM↑ / LPIPS↓). See the text for the details.

2.3　消融实验

为验证本文设计的有效性，重点考察三类关键因素：上采样单元、AVSR先验以及二阶对齐策略。对于上采样单元，我们构建如下变体以评估该单元的有效性：1)采用非学习的插值算子(双线性插值)完成上采样；2)使用本文的超上采样单元，通过预生成上采样卷积核并用于重建。为验证多尺度频率先验是否为重建过程提供有效指导信息，并对比其他先验存在优势，我们对比三种配置：1)不引入先验；2)引入VGG先验；3)引入本文提出的多尺度频率先验。对于二阶对齐策略和前瞻机制的效果验证，我们在相同网络基座上改变对齐分支的启用方式，对比单向与双向对齐在时域信息聚合中的差异：1)聚合历史信息的分支；2)前瞻未来信息的分支。通过改变上述配置所产生的所有变体如表4所示，其结果清晰表明：先验与对齐是带来主要增益的两个环节。具体而言，在固定上采样方式与双向二阶对齐的条件下，引入先验可带来稳定改进：相较于不使用先验，VGG先验平均提升0.13dB，同时LPIPS也整体降低；进一步将VGG先验替换为本文的拉普拉斯金字塔频率先验后，性能继续获得一致的小幅提升，说明多尺度频率约束能够在不同放大倍率下有效提供尺度一致的指导。对于二阶对齐策略，在相同先验与上采样设置下，双向二阶对齐显著优于单向对齐。表明双向对齐能更充分利用相邻帧信息，在运动、遮挡或位移估计误差存在时提升时序聚合的鲁棒性并抑制错位引起的伪影。上采样部分的对比表明：在相同先验与双向二阶对齐下，超网络上采样较插值在各倍率均稳定提升。

为验证前瞻机制中滑动窗口长度L的影响，本文在REDS数据集上对不同L值进行了消融实验，结果如表5所示。可以看出，当L从0增大到2时，各倍率下的定量指标整体持续提升，说明适度引入未来帧信息有助于增强时序信息利用并改善重建质量。

当L进一步增大到3时，PSNR和SSIM仅在部分倍率下有轻微提升，整体收益趋于饱和，而LPIPS在部分倍率下反而略有下降。特别是在×8倍率下，性能甚至出现回落，表明在大倍率重建场景中，远距离未来帧所提供的有效辅助信息有限。考虑到更大的窗口还会带来更高的缓存需求与推理延迟，本文最终采用L=2作为默认设置，以在性能与效率之间取得较好平衡。以上消融验证了本文关键改进对任意倍率视频超分的稳定收益，并为完整模型配置提供了直接的定量依据。

3 结 论

本文提出了一种名为SL-AVSR的任意倍率视频超分辨率方法，该方法结合了二阶可形变对齐与拉普拉斯金字塔先验。在基于前瞻机制的时序建模框架下，通过引入多尺度频率先验引导、二阶可形变对齐以及超网络上采样单元，有效提升了复杂运动场景下的空间细节还原能力和时序一致性，并具备优异的倍率泛化性能。在多个基准数据集和未见退化设置上的实验结果表明，SL-AVSR在PSNR、SSIM和LPIPS等客观指标上达到了整体最优或具有竞争力的性能，同时在主观视觉质量与时序稳定性方面表现更为稳健。消融实验进一步验证了多尺度频率先验、二阶可形变对齐模块以及前瞻式时序框架等关键设计对性能提升的有效贡献。

未来工作将从三个方面展开：进一步提升对齐模块在快速运动与遮挡条件下的鲁棒性，扩大时域建模范围以更充分利用长序列信息，并探索与生成式先验的结合，以提升极端倍率与复杂真实退化场景下的重建质量。

参考文献(References)

Behjati P., Rodriguez P., Mehri A., Hupont I., Tena C. F., & Gonzalez J. (2021). Overnet: Lightweight multi-scale super-resolution with overscaling network. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 2694-2703). [DOI:10.1109/wacv48630.2021.00274]

Burt P. J., & Adelson E. H. (1987). The Laplacian pyramid as a compact image code. In Readings in computer vision (pp. 671-679). Morgan Kaufmann. [DOI: 10.1515/9781400827268.28]

Caballero J., Ledig C., Aitken A., Acosta A., Totz J., Wang Z., & Shi W. (2017). Real-time video super-resolution with spatio-temporal networks and motion compensation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4778-4787). [DOI: 10.1109/cvpr.2017.304]

Cao J., Wang Q., Xian Y., Li Y., Ni B., Pi Z., ... & Van Gool, L. (2023). Ciaosr: Continuous implicit attention-in-attention network for arbitrary-scale image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp.1796-1807). [DOI: 10.1109/cvpr52729.2023.00179]

Chambolle A. (2004). An algorithm for total variation minimization and applications. Journal of Mathematical imaging and vision, 20(1), 89-97.[DOI: 10.1023/b:jmiv.0000011325.36760.1e]

Chan K. C., Wang X., Yu K., Dong C., & Loy C. C. (2021). 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 (pp. 4947-4956). [DOI: 10.1109/cvpr46437.2021.00491]

Chan K. C., Zhou S., Xu X., & Loy C. C. (2022). Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5972-5981). [DOI: 10.1109/cvpr52688.2022.00588]

Chan K. C., Zhou S., Xu X., & Loy C. C. (2022). Investigating tradeoffs in real-world video super-resolution. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp.5962-5971). [DOI: 10.1109/cvpr52688.20 22.00587]

Chen H. W., Xu Y. S., Hong M. F., Tsai Y. M., Kuo H. K., & Lee C. Y. (2023). Cascaded local implicit transformer for arbitrary-scale super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18257-18267). [DOI: 10.1109/cvpr52729.2023.01751]

Chen Y., Liu S., & Wang X. (2021). Learning continuous image representation with local implicit image function. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp.8628-8638).[DOI: 10.1109/cvpr46437.2021.00852]

Chen Y. H., Chen S. C., Lin Y. Y., & Peng W. H. (2023). 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 (pp. 23131-23141). [DOI: 10.1109/iccv51070.2023.02114]

Chen Z., Chen Y., Liu J., Xu X., Goel V., Wang Z., ... & Wang, X. (2022). Videoinr: Learning video implicit neural representation for continuous space-time super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp.2047-2057). [DOI: 10.1109/cvpr52688.2022.00209]

Donoho D L. 2006. Compressed sensing. IEEE Transactions on Information Theory, 52(4): 1289-1306[DOI: 10.1137/1.9781611976120.ch2]

Fu Y., Chen J., Zhang T., & Lin Y. (2021). Residual scale attention

network for arbitrary scale image super-resolution. Neurocomputing, 427, 201-211. [DOI: 10.1016/j.neucom.2020.11.010]

Hornik K., Stinchcombe M., & White H. (1989). Multilayer feedforward networks are universal approximators. Neural networks, 2 (5), 359-366. [DOI: 10.1016/0893-6080(89)90020-8]

Hu J., Xia B., Chen B., Yang W., & Zhang L. (2025, April). Gaussiansr: High fidelity 2d gaussian splatting for arbitrary-scale image super-resolution. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 4, pp. 3554-3562). [DOI: 10.1609/aaai.v39i4.32369]

Hu X., Mu H., Zhang X., Wang Z., Tan T., & Sun J. (2019). Meta-SR: A magnification-arbitrary network for super-resolution. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1575-1584). [DOI: 10.1109/cvpr.2019.00167]

Irani M., & Peleg S. (1991). Improving resolution by image registration. CVGIP: Graphical models and image processing, 53 (3), 231-239. [DOI: 10.1016/1049-9652(91)90045-l]

Jiang J J, Cheng H, Li Z Y, Liu X M and Wang Z Y. 2023. Deep learning based video-related super-resolution technique: a survey. Journal of Image and Graphics, 28(7): 1927-1964 (江俊君, 程豪, 李震宇, 刘贤明, 王中元. 2023. 深度学习视频超分辨率技术综述. 中国图象图形学报, 28(7): 1927-1964) [DOI: 10.11834/jig.220130]

Jin Y T, Song H H and Liu Q S. 2022. Super-resolution video frame reconstruction through lightweight attention constraint alignment network. Journal of Image and Graphics, 27(10): 2984-2993 (靳雨桐, 宋慧慧, 刘青山. 2022. 轻量级注意力约束对齐网络的视频超分重建. 中国图象图形学报, 27(10): 2984-2993) [DOI: 10.11834/jig.210345]

Kim E., Kim H., Jin K. H., & Yoo J. (2025). BF-STVSR: B-Splines and Fourier---Best Friends for High Fidelity Spatial-Temporal Video Super-Resolution. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 28009-28018). [DOI: 10.1109/cvpr52734.2025.02608]

Lai W. S., Huang J. B., Ahuja N., & Yang M. H. (2017). Deep laplacian pyramid networks for fast and accurate super-resolution. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 624-632). [DOI: 10.1109/cvpr.2017.618]

Lee J., & Jin K. H. (2022). Local texture estimator for implicit representation function. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1929-1938). [DOI: 10.1109/cvpr52688.2022.00197]

Li Z., Liu H., Shang F., Liu Y., Wan L., & Feng W. (2024, March). SAVSR: arbitrary-scale video super-resolution via a learned scale-adaptive network. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 4, pp. 3288-3296). [DOI: 10.1609/aaai.v38i4.28114]

Liang J., Cao J., Sun G., Zhang K., Van Gool L., & Timofte R. (2021). Swinir: Image restoration using swin transformer. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1833-1844). [DOI: 10.1109/iccvw54120.2021.00210]

Mairal J., Bach F., & Ponce J. (2014). Sparse modeling for image and vision processing. Foundations and Trends® in Computer Graphics and Vision, 8(2-3), 85-283. [DOI: 10.1561/9781680830095]

Michalkiewicz M., Pontes J. K., Jack D., Baktashmotlagh M., & Eriksson A. (2019). Implicit surface representations as layers in neural networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4743-4752). [DOI: 10.1109/iccv.2019.00484]

Mildenhall B., Srinivasan P. P., Tancik M., Barron J. T., Ramamoorthi R., & Ng R. (2021). Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1), 99-106. [DOI: 10.1109/iccv.2019.00484]

Nah S., Baik S., Hong S., Moon G., Son S., Timofte R., & Mu Lee K. (2019). 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 (pp. 0-0). [DOI: 10.1109/cvprw.2019.00251]

Ranjan A., & Black M. J. (2017). Optical flow estimation using a spatial pyramid network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4161-4170). [DOI: 10.1109/cvpr.2017.291]

Shang W., Ren D., Yang Y., Zhang H., Ma K., & Zuo W. (2023). Joint video multi-frame interpolation and deblurring under unknown exposure time. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13935-13944). [DOI: 10.1109/cvpr52729.2023.01339]

Shang W., Ren D., Zhang W., Fang Y., Zuo W., & Ma K. (2024, September). Arbitrary-Scale Video Super-Resolution with Structural and Textural Priors. In European Conference on Computer Vision (pp. 73-90). Cham: Springer Nature Switzerland. [DOI: 10.1007/978-3-031-72998-0_5]

Shi W., Caballero J., Huszár F., Totz J., Aitken A. P., Bishop R., ... & Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1874-1883). [DOI: 10.1109/cvpr.2016.207]

Sitzmann V., Martel J., Bergman A., Lindell D., & Wetzstein G. (2020). Implicit neural representations with periodic activation functions. Advances in neural information processing systems, 33, 7462-7473. [DOI: 10.48550/arXiv.2006.09661]

Song G., Sun Q., Zhang L., Su R., Shi J., & He Y. (2023). OPE-SR: Orthogonal position encoding for designing a parameter-free upsampling module in arbitrary-scale image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10009-10020). [DOI: 10.1109/cvpr52729.2023.00965]

Tao X., Gao H., Liao R., Wang J., & Jia J. (2017). Detail-revealing deep video super-resolution. In Proceedings of the IEEE international conference on computer vision (pp. 4472-4480). [DOI: 10.1109/iccv.2017.479]

Ulyanov D., Vedaldi A., & Lempitsky V. (2018). Deep image prior. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 9446-9454). [DOI: 10.1109/cvpr.2018.00984]

Wang L., Wang Y., Lin Z., Yang J., An W., & Guo Y. (2021). Learning a single network for scale-arbitrary super-resolution. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 4801-4810). [DOI: 10.1109/iccv48922.2021.00476]

Wang X., Chan K. C., Yu K., Dong C., & Change Loy C. (2019). Edvr: Video restoration with enhanced deformable convolutional networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (pp. 0-0). [DOI: 10.1109/cvprw.2019.00247]

Wang X., Chen X., Ni B., Wang H., Tong Z., & Liu Y. (2023). Deep arbitrary-scale image super-resolution via scale-equivariance pursuit. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1786-1795). [DOI: 10.1109/cvpr52729.2023.00178]

Zhou Y., Dong B., Wu Y., Zhu W., Chen G., & Zhang Y. (2023, August). Dichotomous Image Segmentation with Frequency Priors. In IJCAI (pp. 1822-1830). [DOI: 10.24963/ijcai.2023/202]

作者简介

王志翔，男，硕士研究生，主要研究方向为图像视频修复。E-mail：wzx036@tju.edu.cn

张雅媛，女，本科生，主要研究方向为图像视频修复。E-mail：anna_zhang_0451@tju.edu.cn

尚玮，男，博士研究生，主要研究方向为图像视频修复。E-mail：csweishang@gmail.com

杨柳，女，教授，主要研究方向为机器学习、数据挖掘。E-mail：yangliuyl@tju.edu.cn

朱鹏飞，男，教授，主要研究方向为机器学习、计算机视觉。E-mail：zhupengfei@tju.edu.cn

任冬伟，男，英才副教授，主要研究方向为机器学习、计算机视觉。E-mail：rendw@tju.edu.cn