文章编号：1007-2780(2026)03-0440-12

基于OFCA-Transformer的轻量化视频超分辨率重建

任朋炀[1]，庞　凯[2*]
(1. 合肥工业大学 机械工程学院，安徽 合肥 230002；
2. 东北大学 机械工程与自动化学院，辽宁 沈阳 110167)

摘要：针对现有视频超分辨率方法在复杂运动场景下存在的帧间对齐不准确、时序信息利用不充分，以及传统注意力机制计算复杂度高等问题，本文提出一种融合光流引导交叉注意力的视频超分辨率网络(OFCA-Transformer)。首先，设计一个轻量级的多尺度光流估计模块，生成多粒度运动信息；其次，创新性引入光流引导的交叉注意力机制，以光流预测位置为中心建立局部注意力窗口，实现显式几何先验与隐式内容感知的深度融合，在提升对齐精度的同时显著降低计算复杂度；最后，构建分层特征聚合模块，在Transformer架构内实现更有效的时空特征融合。在放大因子分别为×2、×3、×4时，将本文的研究方法与其他方法在3个公开数据集进行对比。结果表明，OFCA-Transformer在多个数据集上的PSNR值与其他先进方法相比仅差0.16 dB，而模型参数量降低82.8%，有效地提高了计算效率。此外，本文所提的研究方法在复杂运动场景下表现出更精确的细节恢复和更好的时间一致性，客观上在各个放大因子下均取得较好的定量指标。

关　键　词：视频超分辨率；Transformer；光流估计；交叉注意力；运动对齐
中图分类号：TP391.4　　**文献标识码**：A　　**doi**：10.37188/CJLCD.2025-0256　　**CSTR**：32172.14.CJLCD.2025-0256

Lightweight video super-resolution reconstruction based on OFCA-Transformer

REN Pengyang[1], PANG Kai[2*]

(1. *School of Mechanical Engineering*, *Hefei University of Technology*, *Hefei* 230002, *China*;
2. *School of Mechanical Engineering and Automation*, *Northeastern University*, *Shenyang* 110167, *China*)

Abstract: To address the limitations of existing video super-resolution methods in complex motion scenes-including inaccurate frame-to-frame alignment, insufficient utilization of temporal information, and high computational complexity of traditional attention mechanisms, this paper proposes an optical flow-guided cross-attention video super-resolution network (OFCA-Transformer). First, a lightweight multi-scale optical flow estimation module is designed to generate multi-granularity motion information. Second, we innovatively introduce a flow-guided cross-attention mechanism. By establishing local attention windows centered on flow-predicted positions, we achieve an explicit fusion of geometric priors with implicit content awareness. This approach significantly enhances alignment accuracy while substantially reducing computational complexity. Additionally, we construct a hierarchical feature aggregation module to enable more efficient

收稿日期：2025-12-28；**修订日期**：2026-02-02.
*通信联系人，E-mail：neu_pangkai@163.com

spatio-temporal feature fusion within the Transformer architecture. Our method was evaluated against other approaches on three public datasets at magnification factors of ×2, ×3, and ×4. The results demonstrate that OFCA-Transformer achieves PSNR values only 0.16 dB lower than the state-of-the-art methods across multiple datasets, while reducing model parameters by 82.8%, effectively improving computational efficiency. Furthermore, the proposed method exhibits more precise detail recovery and better temporal consistency in complex motion scenes, objectively achieving superior quantitative metrics across all magnification factors.
Key words: video super-resolution; Tranformer; optical flow estimation; cross-attention; feature fusion

1　引　　言

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[1]。超分辨率技术（Super-Resolution，SR）是近年来计算机视觉和图像处理领域中的一个研究热点，其主要目标是将低分辨率图像/视频转换为高分辨率图像/视频[2]。与单幅图像超分辨率（SISR）相比，视频超分辨率（VSR）的核心在于如何有效利用相邻帧间的互补信息，并通过精确的帧间对齐和融合来重构高质量细节。视频超分辨率主要应用于影像修复[3]、网络传输[4]、视频监控[5]和超高清产业[6]等领域。早期的VSR方法直接沿用图像超分领域的方法重建每帧视频，并没有考虑到视频的时间维度，从而导致出现视频帧之间不连贯等问题，重建效果不佳。

近年来，深度学习已主导VSR领域的发展。最早是Kappeler[7]等人提出一种视频超分辨率重建网络（Video Super-Resolution Network，VSRNet），首次将卷积神经网络应用于视频超分辨率重建。2021年，Chan等人提出的BasicVSR[8]和其改进的BasicVSR＋＋[9]都是使用循环结构，通过双向传播充分利用整个视频序列的时序信息，取得了显著进展。同年，受Transformer在自然语言处理和高层视觉任务中成功的启发，研究者开始将其引入VSR领域。VSR-Transformer[10]是这一方向的先驱，它首次将纯Transformer架构应用于VSR，通过时空自注意力机制有效捕捉了全局依赖关系，展现了巨大潜力。2022年，Wang[11]等人提出一种融合光流运动补偿与Swin Transformer的时空视频超分辨率方法（FlowST），通过光流实现帧间精确对齐，并利用分层窗口自注意力高效建模时空依赖关系。光流为视频帧间的运动提供了明确的、像素级的描述，长期以来一直是VSR中运动补偿的关键技术。2024年，夏振平[12]等人在其提出的基于混合时空卷积的轻量级视频超分辨率重建网络中设计了一种基于注意力机制的运动补偿模块，旨在有效减少错误特征融合所带来的负面影响。2024年，林坚普[13]团队引入级联残差机制对Transformer网络结构进行优化，并将其应用于图像超分辨率重建任务。该方法有效增强了卷积神经网络在多尺度特征层面的自适应学习能力，进一步改善了超分辨率算法的整体性能。Xue[14]等人在端到端训练范式下，设计适配视频复原任务的光流估计模块以学习贴合复原特征的光流，并构建了任务导向型光流网络Toflow。2025年，陈清江[15]等人提出了一种多维度聚合Transformer网络，设计了空间-通道交互模块，并将其集成于Transformer层中，使重建效果更加清晰，提升了模型性能。Jin[16]等人基于对光场图像进行多尺度、实时、高分辨率的重建，提出的RTU-Net模型，可应用于各种应用领域，以从微观尺度到宏观尺度深入了解体积成像。贺兴[17]等人设计了一种融入多维注意力网络的单图像超分辨率重建方法，结合通道注意力、自注意力等，提升了对不同尺度特征的捕捉能力，在恢复高分辨率图像时保留更多的细节和更好的全局一致性。阎刚[18]等人引入了全方位状态空间轻量化超分辨率模型，并提出残差全方位空间组作为核心组块，从而实现全方位特征提取。Yang[19]等人提出的一种多尺度时空特征融合的视频超分辨率重建方法（STVSR），通过光流估计对相邻帧进行精确运动建模，并在此基础上实现了多尺度时空特征融合。然而，当存在遮挡或者快速

运动等复杂运动场景时，视频帧依旧存在帧间对齐不准确、时序信息利用不充分等问题。

为解决上述问题，本文在VSR-Transformer的基础上提出了一种全新的融合光流引导交叉注意力的视频超分辨率网络(OFCA-Transformer)，其核心思想是利用光流为Transformer的注意力计算提供空间约束，从而提高显式运动的精确性和隐式注意力的内容自适应性。

本文的主要贡献如下：(1) 提出了光流引导的交叉注意力机制(OFCA)，利用光流矢量定义查询位置的局部注意力窗口，将全局注意力计算转化为高效的局部操作，降低计算复杂度，提高特征对齐的准确性；(2) 设计了轻量级多尺度光流估计模块，以相邻低分辨率(LR)帧为输入，并通过一个轻量的金字塔网络估计双向光流，在引入少量额外参数的条件下为OFCA模块提供可靠的多粒度运动先验；(3) 构建了分层特征聚合模块，用OFCA模块替代部分自注意力层，并设计了分层特征聚合策略，实现更高效的时空信息整合；(4) 进行全面充分的实验验证：在多个公开数据集上，本文所提方法在客观指标和主观视觉对比上均超越了多数现有主流方法，同时保持了合理的模型复杂度，验证了本文方法的有效性、高效性与泛化能力。

2 VSR-Transformer

Transformer架构凭借其自注意力机制，在捕获长距离依赖关系方面展现出强大能力，近年来被成功应用于VSR任务。VSR-Transformer[10]结构如图1所示，作为开创性工作，首次将纯Transformer用于VSR。它由特征提取器、Transformer编码器和重建网络3部分组成。Transformer编码器核心由N个串联的VSR-Transformer Block组成，每个Block内部又包含两个创新的子层：时空卷积自注意力(STCSA)和双向前馈光流网络(BOFF)。在STCSA层中采用卷积代替全连接自注意力，通过局部卷积核提取时空特征，有效利用了视频帧序列中的局部结构信息，并通过理论分析证明了其在局部模式学习上的优势。在BOFF层中，采用双向光流实现跨帧特征的精准对齐与高效传播，通过预训练的SPyNet估计前后向光流，并利用残差网络实现特征融合，增强帧间信息交互与时空一致性。

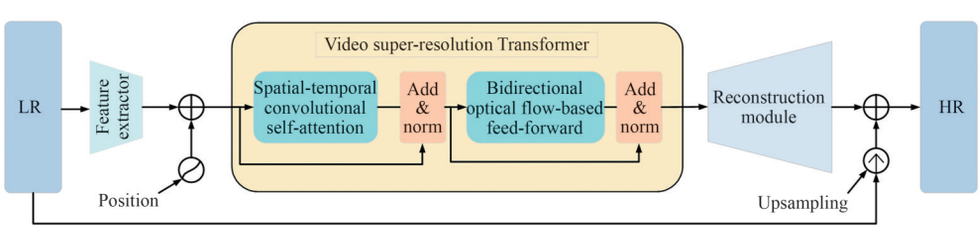


图1 VSR-Transformer结构图

Fig. 1 VSR-Transformer architecture diagram

在VSR-Transformer中，首先，通过由残差块组成的特征提取器将低分辨率(LR)视频帧序列$\{V_1, V_2\cdots, V_T\}$转换为更具有表达力的深度特征图；其次，由Transformer编码器中的STCSA层将来自上一个Block的特征图$\{X_1, X_2, \cdots, X_T\}$生成包含局部和全局信息的特征图$\{X_1', X_2', \cdots X_T'\}$，BOFF层则通过SPyNet对STCSA生成的特征图进行对齐和传播，形成最终的特征图序列$\{X_1'', X_2'', \cdots X_T''\}$。最后，在重建模块中，VSR-Transformer采用多层残差块进行特征细化，并通过上采样模块从增强后的特征中重建出高分辨率(HR)视频帧序列。

3 本文方法

3.1 网络整体架构

本文提出的光流引导交叉注意力机制(OFCA-Transformer)整体结构如图2所示，该模型是基于VSR-Transformer改进而来，包含4个主要组件：特征提取模块、多尺度光流估计模块、分层特征聚合模块和重建模块。首先，本文假设D为视

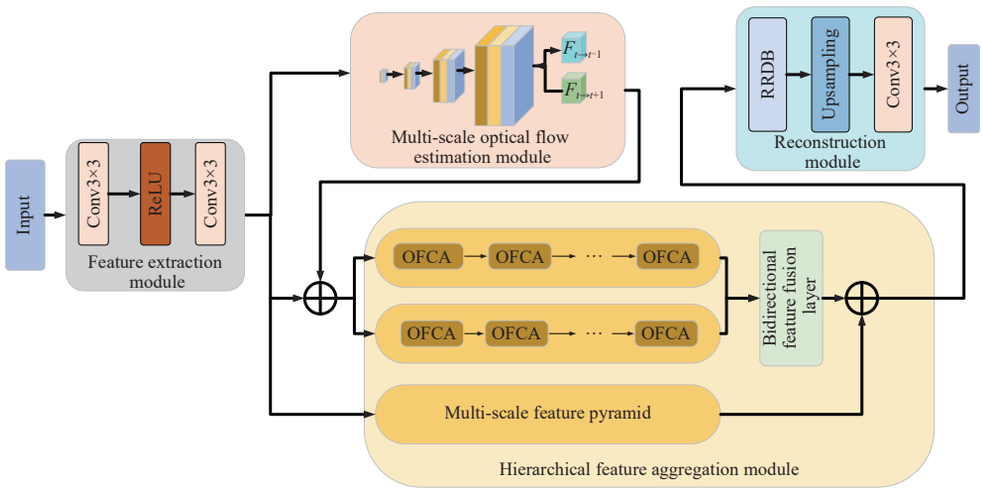


图 2　OFCA-Transformer整体结构图

Fig. 2　Overall architecture of OFCA-Transformer

频的分布，$\{V_1, V_2, \cdots, V_T\} \sim D$ 为低分辨率(LR)视频序列，其中 $V_t \in \boldsymbol{R}^{3\times W\times H}$ 为第 t 个LR帧。使用特征提取模块通过一个共享权重的浅层卷积网络从LR视频帧中学习特征 $\chi=\{X_1, X_2, \cdots X_T\}$，其中 $X_t \in \boldsymbol{R}^{C\times W\times H}$ 是第 t 个特征。利用多尺度光流估计[20]模块提取像素级的运动先验，然后利用分层特征聚合模块双向传播机制分别处理视频序列输出特征信息，最后通过重建模块生成高分辨率中心帧。

3.2　多尺度光流估计模块

本文设计了多尺度光流估计模块，如图3所示。该模块通过轻量化金字塔网络[21](包含1/8、1/4、1/2及原尺度4个层级)实现相邻帧间像素级运动向量的精准估计。该模块以“由粗到精”为“递归细化核心策略”：在(1/8)粗尺度层面直接预测大范围运动，如公式(1)所示：

$$\begin{cases} \text{flow}_4 = f_{\theta_4}(\beta) \\ \beta = \text{concat}\left(F_t^4, F_{t\pm1}^4, \text{corr}\left(F_t^4, F_{t\pm1}^4\right)\right) \end{cases}, \quad (1)$$

式中：F_t^4、$F_{t\pm1}^4$ 为第四层特征图，$\text{corr}\left(F_t^4, F_{t\pm1}^4\right)$ 为相关性计算，f_{θ_4} 为第四层预测网络。在各精细层，采用“上采样＋残差修正”策略。首先将粗尺

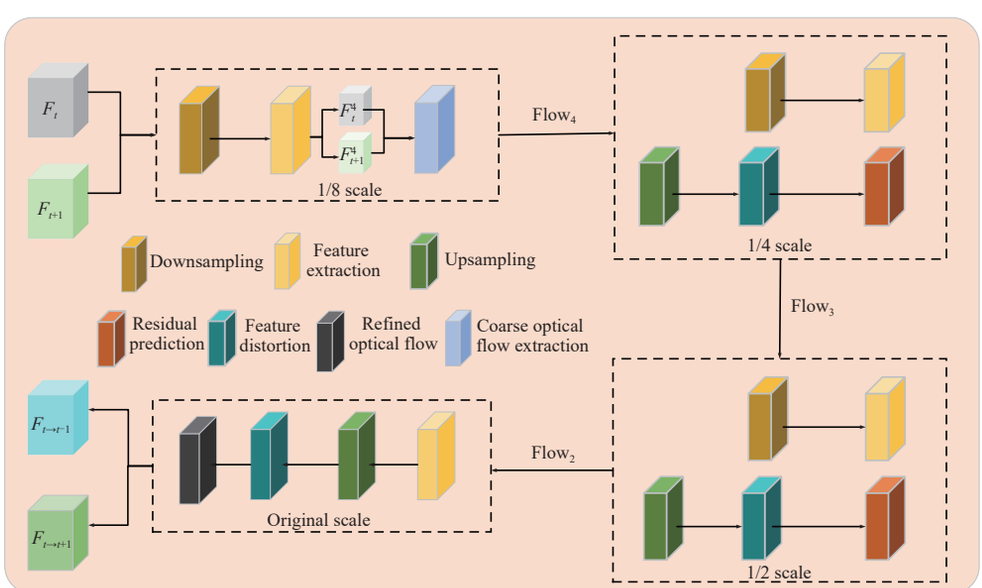


图 3　多尺度光流估计模块

Fig. 3　Multi-scale optical flow estimation module

度得到的 flow_4 进行上采样处理，其次利用上采样后的光流对相邻帧特征进行扭曲对齐，最后通过残差学习策略预测细节特征，实现特征的精准补全与优化。其过程如公式(2)和公式(3)所示：

$$\text{flow}_{\text{up}}^{t}=\text{upsample}\left(\text{flow}^{t+1}\right)\times 2\,, \tag{2}$$

$$\text{flow}_{t+1,\text{warped}}^{t}=\text{warp}\left(F_{t+1}^{t},\text{flow}_{\text{up}}^{u}\right), \tag{3}$$

其中扭曲操作使用可微分的双线性采样，如公式(4)～(7)所示：

$$\text{flow}_{t+1,\text{warped}}^{t}(p)=\sum_{\varphi}F_{t\pm1}^{t}(q)\chi\eta\,, \tag{4}$$

$$\varphi=q\in N\left(p+\text{flow}_{\text{up}}^{t}(p)\right), \tag{5}$$

$$\chi=\left(1-\left|p_x-q_x\right|\right), \tag{6}$$

$$\eta=\left(1-\left|p_y-q_y\right|\right). \tag{7}$$

最终输出双向光流场 $F_{t\to t-1}$ 和 $F_{t\to t+1}$，为后面光流引导的交叉注意力模块提供显示运动先验，将全局注意力计算约束于以光流预测位置为中心的局部窗口内，不仅将计算复杂度从 $O((HW)^2)$ 显著优化至 $O(HW\times S^2)$，更有效解决了复杂运动场景下的帧间对齐问题。

3.3 OFCA 模块

为有效实现视频序列的精准对齐和高质量特征融合，本文设计了光流引导交叉注意力模块(OFCA)，如图 4 所示，该模块构成了 OFCA-Transformer 的核心对齐与融合单元。OFCA 模块集成了显式运动估计和隐式注意力机制[22]的协同优势，通过端到端的可学习架构实现时空特征的有效聚合。OFCA 模块分为两个部分：显式运动补偿和隐式特征融合。在显式运动补偿阶段，通过轻量级光流网络计算相邻帧与参考帧之间的双向光流场，利用得到的运动矢量对相邻帧特征进行空间变换和初步对齐。在隐式特征融合阶段，采用光流引导的局部交叉注意力机制，以光流预测的对应位置为中心建立动态注意力窗口，在保持计算效率的同时实现精细的特征重校准。

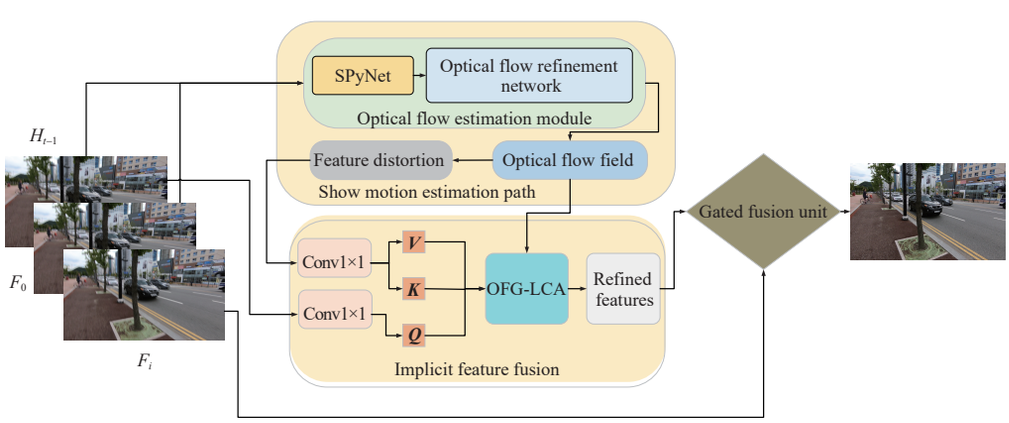


图 4 OFCA 模块结构

Fig. 4 OFCA module structure

OFCA 模块的计算过程如下：给定参考帧特征 F_0 和相邻帧特征 F_i，模块首先利用一个轻量级光流网络 f_{flow} 计算相邻帧到参考帧的稠密光流场，如公式(8)所示：

$$\text{OF}_{i\to 0}=f_{\text{flow}}(I_i,I_0)\,. \tag{8}$$

利用该光流场，对输入特征进行初步的空间变换和对齐，如公式(9)所示：

$$F_i^{\omega}=\omega\left(F_i,\text{OF}_{i\to 0}\right), \tag{9}$$

其中：ω 表示可微的双线性采样(扭曲)操作。

在隐式特征融合阶段，模块通过交叉注意力机制进行精细的特征重校准。首先，通过线性投影生成查询(Query，$\boldsymbol{Q}$)、键(Key，$\boldsymbol{K}$)和值(Value，$\boldsymbol{V}$)向量，如公式(10)所示：

$$\begin{cases}\boldsymbol{Q}=W_qF_0\\ \boldsymbol{K}=W_kF_i^{\omega}\\ \boldsymbol{V}=W_vF_i^{\omega}\end{cases}, \tag{10}$$

式中 W_q、、W_k、、W_v 为可学习的投影权重。

本文的核心在于引入光流引导的局部注意力机制。对于参考帧上的任一空间位置 p，利用光流 $\text{OF}_{i\to 0}(p)$ 确定其在相邻帧中的粗略对应点 p'。以 p' 为中心，定义一个固定大小为 $S\times S$ 的局部窗口 $\Omega(p')$。注意力权重的计算被限制在该局

部窗口内，如公式(11)所示：

$$\alpha_{p,q}=\frac{e^{\frac{\langle Q(p),W(q)\rangle}{\sqrt{d}}}}{\sum_{q'\in\Omega(q')}e^{\frac{\langle Q(p),W(q)\rangle}{\sqrt{d}}}},\forall q\in\Omega(p'),\quad(11)$$

式中：d为通道维数，$\langle\cdot,\cdot\rangle$表示点积运算。精炼后的输出特征通过加权求和得到功能公式(12)：

$$O_{(p)}=\sum_{q\in\Omega(p')}\alpha_{p,q}\cdot V(q).\quad(12)$$

最后，通过一个门控融合单元自适应地整合精炼特征与原始参考帧特征，生成模块的最终输出H_t，如公式(13)～(14)所示：

$$G=\sigma\left(W_g[O,F_0]\right),\quad(13)$$

$$H_t=G\odot O+(1-G)\odot F_0,\quad(14)$$

其中：σ为Sigmoid函数，W_g为卷积权重，$[\cdot,\cdot]$表示通道拼接，$\odot$表示逐元素相乘。本模块通过光流先验缩小注意力范围，实现计算高效性，显著降低计算复杂度；在保持精确对齐的同时，融合注意力机制的自适应特性，确保对齐的精确性；并且对光流估计误差具有一定容错能力，可通过注意力权重自动调整，展现出强鲁棒性。

3.4　分层特征聚合模块

为充分挖掘视频序列中的时空上下文信息，本文设计了双向传播特征融合机制，如图5所示。该模块由多个堆叠的Transformer块组成，每个块内部由多个OFCA模块构成，以替代部分标准自注意力层。该架构通过双向传播流程捕获时序依赖性，并借助多尺度金字塔融合不同层次的视觉特征。在双向传播路径中，网络分别沿前向与后向两个独立分支处理特征序列。前向传播流按时间顺序聚合历史信息，后向传播流逆序整合未来上下文，其计算过程如公式(15)所示：

$$\begin{cases}H_i^{\text{for}}=f_{\text{att}}\left(H_{i-1}^{\text{for}},F_i,\text{OF}_{i\to t}\right)\\H_i^{\text{back}}=f_{\text{att}}\left(H_{i+1}^{\text{back}},F_i,\text{OF}_{i\to t}\right)\end{cases},\quad(15)$$

式中：H_i^{for}、H_{i-1}^{for}和H_i^{back}、H_{i+1}^{back}分别表示前向和后向传播在i时刻的隐藏状态，f_{att}为基于注意力机制的特征变换函数，$\text{OF}_{i\to t}$为相邻帧到参考光流的光流场。

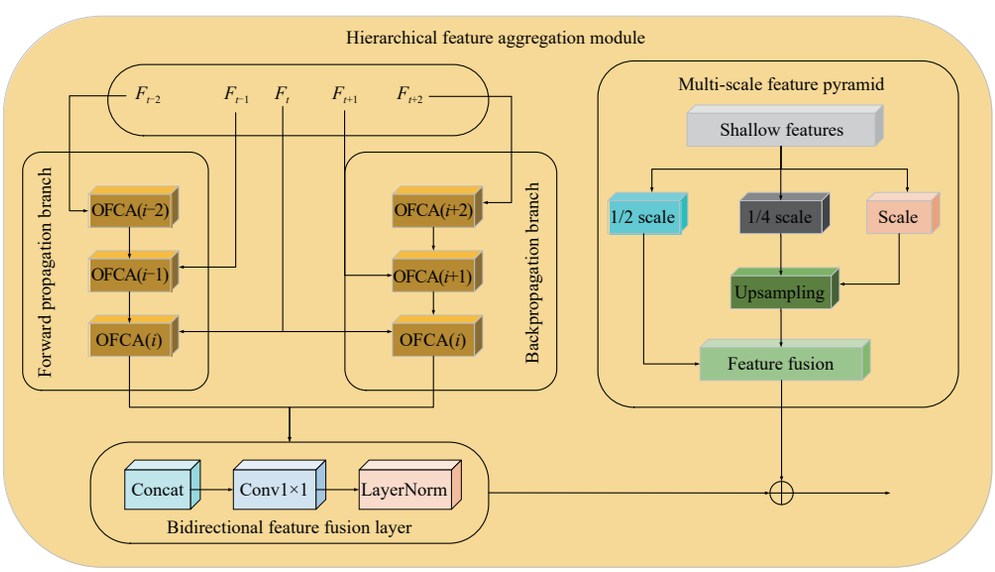


图5　分层特征聚合模块结构

Fig. 5　Structure of the hierarchical feature aggregation module

本研究还构建了三级特征金字塔用于多尺度特征提取。首先，用原尺度分支来保持原始空间分辨率，专注于细粒度局部特征的提取；其次，1/2尺度分支通过步长2卷积实现特征下采样，在扩大感受野的同时保持结构信息；最后，通过1/4尺度分支进一步下采样至原分辨率1/4，重点捕获全局语义上下文信息。各尺度特征通过精心设计的融合路径进行整合。首先将1/2尺度与1/4尺度特征分别上采样至原分辨率，随后通过跳跃连接与原始尺度特征在通道维度拼接，最后利用1×1卷积实现跨尺度特征重校准。

3.5 损失函数

本文采用复合损失函数来平衡重建图像的像素级精度与视觉感知质量，其总体目标函数定义如公式(16)所示：

$$L=\lambda_1 L_{\text{char}}+\lambda_2 L_{\text{per}}\ , \tag{16}$$

式中：L_{char}为Charbonnier损失[23]，L_{per}为感知损失[24]。通过系统的消融实验，确定各损失项的最优权重系数为$\lambda_1=1$，$\lambda_2=0.1$。Charbonnier损失函数作为基础重建损失项，其数学表达式如公式(17)所示：

$$L_{\text{char}}=\sqrt{\left\| I_i^{\text{SR}}-I_i^{\text{GT}}\right\|^2+\varepsilon^2}\ , \tag{17}$$

式中：I_i^{SR}、I_i^{GT}分别表示生成的第i张低分辨率图像和第i张高分辨率图像的真实值；ε为1×10^{-3}，确保了函数在梯度计算时的数值稳定性，同时赋予损失函数对异常值的鲁棒性。感知损失通过比较深层特征空间中的差异来优化视觉质量。本文采用预训练VGG16网络[25]作为特征提取器，选用多个ReLU层的输出作为深层特征表示，如公式(18)所示：

$$L_{\text{per}}=\left\| \phi_l(I^{\text{HR}})-\phi_l(I^{\text{LR}})\right\|_1\ , \tag{18}$$

其中，ϕ_l为特征映射函数，将输入图像像映射到预训练VGG16网络的某一个ReLU层，输出对应的特征图。感知损失项通过约束重建图像与真实图像在深层特征空间中的相似性，有效提升了重建结果的视觉感知质量。

4 实验结果与分析

4.1 实验环境

本次实验的硬件平台采用Ubuntu 20.04操作系统，搭载Intel®Xeon(R) Gold 6248 CPU@2.50 GHz中央处理器，并采用Tesla A100显卡进行加速计算，显存为80 GB。软件环境配置CUDA 11.6并行计算架构，深度学习框架选用PyTorch 1.13.0，编程语言为Python3.9。

4.2 训练集与数据集

本文使用的训练数据集为公开的视频超分辨率数据集REDS[26]、Vimeo-90K[14]和Vid4[27]。REDS数据集包含300个高质量视频片段，涵盖丰富的真实场景和复杂运动模式，其中240个片段用于训练，30个用于验证，其余30个用于测试。该数据集以其高动态范围和复杂时空特性，成为视频超分辨率研究中的重要基准[26]。Vimeo-90K数据集由64 612个训练序列和7 824个测试序列组成，每个序列包含7个连续帧。该数据集场景多样、内容丰富，已被广泛认可并应用于各类视频相关任务的研究，包括视频超分辨率和视频插值等[14]。Vid4数据集作为经典测试基准，包含4个具有复杂相机运动和丰富纹理细节的黑白视频序列。该数据集虽然规模较小，但其对模型重建质量和时序一致性的严苛要求，使其成为评估视频超分辨率性能的重要标准[27]。

在数据预处理阶段，对视频帧序列采用4倍双三次下采样以获得相应的低分辨率输入。为模拟真实世界的成像退化过程，本文采用标准差为1.6的高斯核对原始高清帧实现图像退化，并附加随机噪声以构建更符合实际应用的训练样本。

4.3 主观视觉对比实验

本文在REDS数据集(×4超分辨率)上，将所提出的OFCA-Transformer与现有的先进方法进行主观视觉对比，包括BasicVSR、VSR-Diff、STVSR、EDVR和VRT，结果如图6～8所示。为尽可能保证对比的公平性，各方法在相同的数据与退化设置下进行训练与测试，并统一采用3帧输入的设置以生成高分辨率视频帧。在相同输入信息量条件下，对不同方法的重建结果进行对比分析。

在×4放大因子下，从视觉细节对比可以观察到：Bicubic插值由于缺乏有效的运动建模与特征学习，重建结果存在边缘模糊与细节缺失等问题。BasicVSR与VSR-Diff相比插值方法能够恢复更多细节纹理信息，但在快速运动时仍会出现运动模糊、重影和伪影等现象，导致细节纹理连续性不足。VRT在整体清晰度方面表现较强，但部分场景中会出现纹理过于平滑的现象，在快速运动场景下还可能产生局部纹理形变。相比之下，OFCA-Transformer在复杂运动(如车牌字符)、细节(如马尾)以及复杂整体场景(如街景)呈现出更清晰的边缘结构与更连贯的纹理细节，重影与伪影现象相对较少。这是因为光流引导的局部交叉注意力机制能够沿预测运动轨迹进行局部化特征聚合，从而降低对齐误差带来的细节破坏；同时，多尺度特征融合为复杂纹理恢复提供了更充分的特征信息支持。



图6　不同方法的复杂运动对比图

Fig. 6　Complex motion comparison chart for different methods

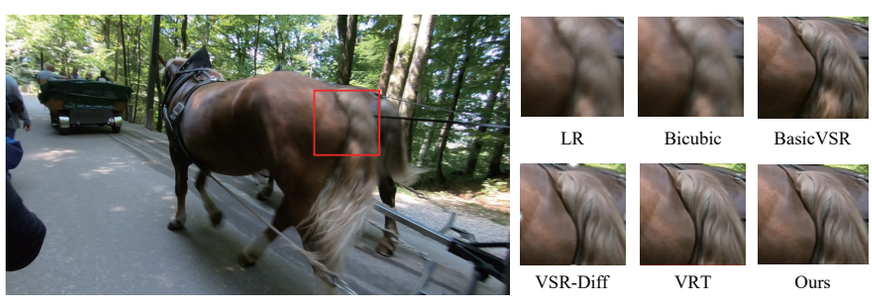


图7　不同方法的细节对比图

Fig. 7　Detailed comparison chart for different methods



图8　不同方法的复杂整体对比图

Fig. 8　Complex overall comparison chart for different methods

4.4　客观对比实验

表1～3分别展示了不同方法在REDS4、Vid4和Vimeo-90K-T数据集上的×2、×3、×4放大因子下的峰值信噪比(PSNR)和结构相似性

表1　REDS4数据集中不同方法的PSNR和SSIM

Tab. 1　PSNR and SSIM of different methods on the REDS4 dataset

方法	×2放大因子		×3放大因子		×4放大因子	
	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM
Bicubic	31.02	0.880 4	28.51	0.810 2	26.95	0.745 6
EDVR	35.23	0.933 1	31.82	0.880 5	29.28	0.865 4
BasicVSR	36.18	0.941 2	32.35	0.893 4	30.12	0.855 0
BasicVSR＋＋	37.01	0.948 3	33.12	0.901 6	31.42	0.875 9
VSR-Diff	38.15	0.949 8	34.00	0.908 9	32.08	0.889 5
VRT	**38.45**	**0.952 8**	**34.23**	**0.912 5**	**32.21**	**0.892 3**
STVSR	38.18	0.950 1	33.95	0.908 0	31.82	0.883 1
OFCA-Net	*38.32*	*0.951 5*	*34.10*	*0.910 8*	*32.05*	*0.889 7*

表 2 Vid4 数据集中不同方法的 PSNR 和 SSIM

Tab. 2 PSNR and SSIM of different methods on the Vid4 dataset

方法	×2 放大因子		×3 放大因子		×4 放大因子	
	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM
Bicubic	24.18	0.698 2	22.45	0.612 3	21.32	0.543 1
EDVR	26.95	0.815 6	24.83	0.734 5	23.41	0.682 3
BasicVSR	27.68	0.832 1	25.42	0.756 8	24.03	0.701 5
BasicVSR++	28.23	0.845 3	26.01	0.773 2	24.65	0.723 4
VSR-Diff	28.88	0.857 6	26.68	0.787 9	25.28	0.743 5
VRT	**29.12**	**0.863 5**	**26.95**	**0.795 4**	**25.48**	**0.752 1**
STVSR	28.75	0.854 0	26.50	0.782 1	25.05	0.738 9
OFCA-Net	*28.96*	*0.859 8*	*26.78*	*0.789 3*	*25.32*	*0.745 8*

表 3 Vimeo-90K-T 数据集中不同方法的 PSNR 和 SSIM

Tab. 3 PSNR and SSIM of different methods on the Vimeo-90K-T dataset

方法	×2 放大因子		×3 放大因子		×4 放大因子	
	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM
Bicubic	33.25	0.912 3	30.18	0.845 6	28.42	0.783 4
EDVR	37.45	0.952 1	34.12	0.901 2	31.85	0.863 4
BasicVSR	38.23	0.958 3	34.95	0.912 3	32.78	0.878 9
BasicVSR++	39.02	0.963 4	35.68	0.923 4	33.45	0.892 3
VSR-Diff	39.80	0.967 9	36.48	0.936 8	34.28	0.906 8
VRT	**39.75**	**0.972 3**	**36.82**	**0.942 3**	**34.56**	**0.912 5**
STVSR	39.65	0.965 5	36.30	0.933 0	33.79	0.952 7
OFCA-Net	**39.98**	**0.969 8**	*36.65*	*0.938 7*	*34.38*	*0.908 2*

(SSIM)定量比较结果。表中,以粗体标注的数字为最优数据,斜体标注的数字为次优数据。实验结果表明,OFCA-Transformer在3个基准数据集的×2、×3、×4多个放大因子上均取得了与最先进方法相媲美的性能。与VRT的性能对比,在REDS4数据集上,×4放大因子中OFCA-Transformer的PSNR达到32.05 dB,与VRT的32.21 dB相比仅差0.16 dB;在Vid4数据集上,×4放大因子中OFCA-Transformer的PSNR达到25.32 dB,与VRT的25.48 dB相比仅差0.16 dB;在Vimeo-90K-T数据集上,×4放大因子中OFCA-Transformer的PSNR达到34.38 dB,与VRT的34.56 dB相比仅差0.18 dB。在推理效率方面,OFCA-Transformer展现出明显优势,处理速度达到VRT的4倍以上。从放大因子来看,在×2、×3、×4三个放大因子上,OFCA-Transformer均表现出稳定的性能。随着放大因子的增加,性能下降趋势与最优方法保持一致。在更富挑战性的×4超分任务中,性能差距在0.2 dB以内。

4.5 模型参数量和运行时间分析

如图9所示,OFCA-Transformer成功实现了高性能与参数高效压缩的兼顾,其总参数量仅为18.2M,占VRT模型(105.7M)参数量的17.2%,

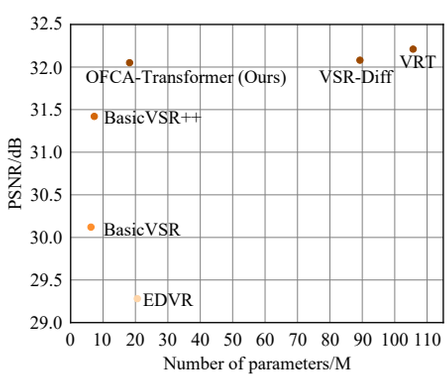


图9 在REDS数据集上放大因子为4时,不同方法的参数量比较。

Fig. 9 Comparison of the number of parameters for different methods at a magnification factor of 4 on the REDS dataset

轻量化优势显著。与其他方法相比，OFCA-Transformer的参数规模处于BasicVSR++(7.3M)与EDVR(20.6M)之间，这一参数量级印证了其在模型复杂度与重建性能间达成了良好的平衡。该模型的轻量化特性主要依托两大核心机制实现：其一为光流引导的局部交叉注意力机制，该机制摒弃了传统全局注意力范式，转而构建以光流预测位置为核心的动态交互窗口，有效消除了大量计算冗余；其二是轻量化多尺度光流估计模块，其采用金字塔网络架构，能够以有限的参数开销完成像素级运动信息的精准提取。

为了进一步突出轻量化优势，本文在REDS4数据集(×4)上补充了参数量(Params)、浮点运算数(FLOPs)及推理时间的对比，如表4所示。可以看出，OFCA-Transformer与STVSR的参数量和FLOPs处于同一水平，但本文方法取得较高的PSNR(超出0.23 dB)和更快的推理速度。这印证了前文分析：OFCA模块通过光流引导实现局部精细化计算，在控制复杂度的同时更有效地提升性能。综合表1～表4，OFCA-Transformer在性能上接近最优的VRT，其效率显著优于VRT，与STVSR相当，在3个数据集上均表现稳定。

表4　不同方法的计算效率对比

Tab. 4　Computational efficiency comparison for different methods

方法	PSNR/dB	SSIM	参数量/M	FLOPs/T	推理时间/(ms·frame^{-1})
BasicVSR++	31.42	0.875 9	**7.3**	**0.73**	32.1
VRT	**32.21**	**0.892 3**	105.7	4.85	88.6
STVSR	31.82	0.883 1	18.5	*1.35*	*24.8*
Ours	*32.05*	*0.889 7*	*18.2*	1.38	**21.5**

4.6　消融实验

为验证本文模型中各模块的有效性，本文设计在REDS4验证集上进行消融实验，结果如表5示。本文设计5组对比实验，分别分析OFCA模块、注意力窗口、分层特征聚合模块和多尺度金字塔对模型性能的影响。实验结果表明：实验1作为最基本的模型，移除了本文提出的所有创新模块，仅保留基于预训练SPyNet的简化单向光流对齐与特征融合框架，其PSNR值仅为31.30 dB，比完整模型小0.75 dB，表明仅依赖传统光流对齐难以充分挖掘跨帧时序信息。实验2在实验1的基础上引入局部窗口自注意力机制，其PSNR值提升了0.2 dB，证明局部注意力有助于建模空间邻域内的特征相关性，并在一定程度上缓解了特征融合不足的问题。实验3将实验2中的局部注意力替换为光流引导的交叉注意力(OFCA)，利用估计的光流动态定位注意力窗口中心，使模型能够沿着运动轨迹进行特征查询与融合，其PSNR值提升到31.80 dB，表明将显式运动先验与隐式注意力机制相结合，有助于提升帧间对齐精度并增强时序特征建模能力，更适用于存在复杂运动的场景。实验4在实验3的基础上进一步设计分层特征聚合物模块，通过并行提取并融合原尺度、1/2尺度及1/4尺度下的特征，实现局部细节与全局语义上下文的高效利用，其PSNR值增加0.2 dB，验证了分层特征聚合模块的有效性。实验5在分层特征聚合模块中引入多尺度特征金字塔结构，增强模型对多尺度信息和复杂纹

表5　消融实验

Tab. 5　Ablation experiments

模型	OFCA模块	注意力窗口	分层特征聚合模块	多尺度特征金字塔	PSNR/dB	SSIM	参数量/M	推理速度/(ms·frame^{-1})
1	×	×	×	×	31.30	0.880 1	15.8	28.3
2	×	局部	×	×	31.50	0.882 0	18.2	25.7
3	√	S=7	×	×	31.80	0.885 5	18.2	24.2
4	√	S=11	√	×	32.00	0.888 0	18.2	21.5
5	√	S=11	√	√	32.05	0.889 7	18.2	21.5

理的捕获重建能力，虽然在整体上PSNR值仅提升0.05 dB，但其能够增强复杂场景下特征表达的稳定性与鲁棒性，且计算成本相当，与光流引导注意力机制互补。

5 结 论

本文提出一种光流引导交叉注意力网络模型（OFCA-Transformer），解决复杂运动场景下存在的帧间对齐不准确、时序信息利用不充分的问题。本文通过引入轻量化多尺度光流估计模块与光流引导的局部交叉注意力机制，既保证了对齐精度又有效降低了模型参数量，并结合分层特征聚合模块实现了多尺度时空特征的高效融合。实验结果表明，在REDS4、Vid4和Vimeo-90K-T等多个基准数据集上，OFCA-Transformer在×2、×3和×4放大因子下的峰值信噪比与现有先进方法相比仅存在约0.16 dB的差距，而模型参数量相较于对比方法减少约82.8%，在显著提升推理效率的同时保持了良好的重建质量与时间一致性。

为了进一步完善评价体系与研究内容，本研究需在后续研究中引入时序感知图像相似度（Tlpips）和时序光流误差（tOF）等时序一致性评价指标，进而从多维度、更全面地验证所提方法在复杂动态运动场景下的有效性与鲁棒性。

参 考 文 献：

[1] 唐麒，赵耀，刘美琴，等. 基于深度学习的视频超分辨率重建算法进展[J]. 自动化学报，2025，51(7)：1480-1524.
TANG Q, ZHAO Y, LIU M Q, *et al*. A review of video super-resolution algorithms based on deep learning [J]. *Acta Automatica Sinica*, 2025, 51(7): 1480-1524. (in Chinese)

[2] 江俊君，程豪，李震宇，等. 深度学习视频超分辨率技术综述[J]. 中国图象图形学报，2023，28(7)：1927-1964.
JIANG J J, CHENG H, LI Z Y, *et al*. Deep learning based video-related super-resolution technique: a survey [J]. *Journal of Image and Graphics*, 2023, 28(7): 1927-1964. (in Chinese)

[3] WAN Z Y, ZHANG B, CHEN D D, *et al*. Bringing old films back to life [C]//*Proceedings of* 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). New Orleans, USA: IEEE, 2022: 17673-17682.

[4] LI G, JI J, QIN M H, *et al*. Towards high-quality and efficient video super-resolution via spatial-temporal data overfitting [C]//*Proceedings of* 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Vancouver, Canada: IEEE, 2023: 10259-10269.

[5] KHAN M A, JAVED K, KHAN S A, *et al*. Human action recognition using fusion of multiview and deep features: an application to video surveillance [J]. *Multimedia Tools and Applications*, 2024, 83(5): 14885-14911.

[6] LUO L G, YI B S, WANG Z Y, *et al*. Efficient lightweight network for video super-resolution [J]. *Neural Computing and Applications*, 2024, 36(2): 883-896.

[7] SUN X, LONG X, HE D L, *et al*. VSRNet: end-to-end video segment retrieval with text query [J]. *Pattern Recognition*, 2021, 119: 108027.

[8] CHAN K C K, WANG X T, YU K, *et al*. BasicVSR: the search for essential components in video super-resolution and beyond [C]//*Proceedings of* 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Nashville, USA: IEEE, 2021: 4947-4956.

[9] CHAN K C K, ZHOU S C, XU X Y, *et al*. BasicVSR++: IMPROVING video super-resolution with enhanced propagation and alignment [C]//*Proceedings of* 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans: IEEE, 2022: 5972-5981.

[10] CAO J Z, LI Y W, ZHANG K, *et al*. Video super-resolution transformer [J/OL]. *arXiv*, 2021: 2106.06847.

[11] WANG X, WANG H, ZHANG M L, *et al*. Combining optical flow and Swin Transformer for space-time video super-resolution [J]. *Engineering Applications of Artificial Intelligence*, 2024, 137: 109227.

[12] 夏振平，陈豪，张宇宁，等. 基于混合时空卷积的轻量级视频超分辨率重建[J]. 光学 精密工程，2024，32(16)：2564-2576.
XIA Z P, CHEN H, ZHANG Y N, *et al*. Lightweight video super-resolution based on hybrid spatio-temporal convolution [J]. *Optics and Precision Engineering*, 2024, 32(16): 2564-2576. (in Chinese)

[13] 林坚普,吴镇城,王崑赋,等. 级联残差优化Transformer网络的图像超分辨率重建[J]. 光学 精密工程,2024,32(12):1902-1914.
LIN J P, WU Z C, WANG K F, *et al*. Cascade residual-optimized image super-resolution reconstruction in Transformer network [J]. *Optics and Precision Engineering*, 2024, 32(12): 1902-1914. (in Chinese)

[14] XUE T F, CHEN B A, WU J J, *et al*. Video enhancement with task-oriented flow [J]. *International Journal of Computer Vision*, 2019, 127(8): 1106-1125.

[15] 陈清江,陈鹏民. 多维度聚合Transformer的图像超分辨率重建[J]. 光学 精密工程,2025,33(12):1955-1970.
CHEN Q J, CHEN P M. MDAT: multi-dimensional aggregation Transformer for image super-resolution reconstruction [J]. *Optics and Precision Engineering*, 2025, 33(12): 1955-1970. (in Chinese)

[16] JIN B Y, HOU Y W, XI P. Bridging microscale and macroscale light-field image reconstruction using 'real-time and universal network' [J]. *Light: Advanced Manufacturing*, 2025, 6(3): 399-401.

[17] 贺兴,王磊,张鹏超,等. 基于多维注意力网络的图像超分辨率重建[J]. 液晶与显示,2025,40(7):1056-1066.
HE X, WANG L, ZHANG P C, *et al*. Image super-resolution reconstruction based on multidimensional attention network [J]. *Chinese Journal of Liquid Crystals and Displays*, 2025, 40(7): 1056-1066. (in Chinese)

[18] 阎刚,宋子怡,耿树泽. 基于全方位状态空间模型的轻量化图像超分辨率重建[J]. 液晶与显示,2025,40(4):642-654.
YAN G, SANG Z Y, GENG S Z. PMambaIR: panoramic vision state space model for lightweight image super-resolution [J]. *Chinese Journal of Liquid Crystals and Displays*, 2025, 40(4): 642-654. (in Chinese)

[19] YANG M Q, WANG Y H, YANG Y, *et al*. Multi-scale spatiotemporal feature fusion for super-resolution video reconstruction in dynamic scenes [J]. *Engineering Applications of Artificial Intelligence*, 2025, 161: 112327.

[20] SUN D Q, YANG X D, LIU M Y, *et al*. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume [C]//*Proceedings of* 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City: IEEE, 2018: 8934-8943.

[21] RANJAN A, BLACK M J. Optical flow estimation using a spatial pyramid network [C]//*Proceedings of* 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). Honolulu: IEEE, 2017: 2720-2729.

[22] TIAN Y P, ZHANG Y L, FU Y, *et al*. TDAN: temporally-deformable alignment network for video super-resolution [C]//*Proceedings of* 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Seattle: IEEE, 2020: 3357-3366.

[23] WANG X L, GIRSHICK R, GUPTA A, *et al*. Non-local neural networks [C]//*Proceedings of* 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City: IEEE, 2018: 7794-7803.

[24] JOHNSON J, ALAHI A, FEI-FEI L. Perceptual losses for real-time style transfer and super-resolution [C]//*Proceedings of the* 14*th European Conference on Computer Vision*. Amsterdam: Springer, 2016: 694-711.

[25] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [C]//*Proceedings of the* 3*rd International Conference on Learning Representations*. San Diego: OpenReview.net, 2015: 1-14.

[26] NAH S, BAIK S, HONG S, *et al*. NTIRE 2019 challenge on video deblurring and super-resolution: dataset and study [C]//*Proceedings of* 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*). Long Beach: IEEE, 2019: 1996-2005.

[27] LIU C, SUN D Q. On Bayesian adaptive video super resolution [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, 36(2): 346-360.

作者简介:



庞　凯,男,硕士研究生,2020年于燕山大学获得学士学位,主要从事视觉处理与图像识别的研究。
E-mail: neu_pangkai@163.com