文章编号　1004-924X(2024)16-2564-13

基于混合时空卷积的轻量级视频超分辨率重建

夏振平[1,3*]，陈　豪[1]，张宇宁[2,4]，程　成[1,3]，胡伏原[1,3]
(1. 苏州科技大学 电子与信息工程学院，江苏 苏州 215009；
2. 东南大学 电子科学与工程学院 显示技术研究中心，江苏 南京 210096；
3. 江苏省工业智能低碳技术工程中心，江苏 苏州 215009；
4. 新型显示与视觉感知石城实验室，江苏 南京 210013)

摘要：针对三维卷积神经网络在视频超分辨率任务上具有较高的计算复杂度以及提取时空特征有限的问题，本文设计了一种基于混合时空卷积的轻量级视频超分辨率重建网络。首先，提出了一个基于混合时空卷积的模块，实现了网络时空特征提取能力的提升以及计算复杂度的降低；其次，提出了一个基于相似性的选择性特征融合模块，进一步增强了相关特征的提取能力；最后，设计了一种基于注意力机制的运动补偿模块，在一定程度上减轻了错误的特征融合的影响。实验结果表明：所提网络可以在视频超分辨率性能和网络复杂度之间取得很好的平衡，而且在基准数据集 SPMCS-11 上 4 倍超分辨率达到 8 frame/s。所提网络满足了边缘设备推理运行中快速、准确等要求。

关　键　词：视频超分辨率；深度学习；三维卷积神经网络；特征融合
中图分类号：TP391.41　　**文献标识码**：A　　**doi**：10.37188/OPE.20243216.2564

Lightweight video super-resolution based on hybrid spatio-temporal convolution

XIA Zhenping[1,3*], CHEN Hao[1], ZHANG Yuning[2,4], CHENG Cheng[1,3], HU Fuyuan[1,3]

(1. *School of Electronic & Information Engineering, Suzhou University of Science and Technology, Suzhou* 215009, *China*;
2. *Display R&D Centre, School of Electronic Science & Engineering, Southeast University, Nanjing* 210096, *China*;
3. *Jiangsu Industrial Intelligent and Low-carbon Technology Engineering Center, Suzhou* 215009, *China*;
4. *Shi-Cheng Laboratory for Information Display and Visualization, Nanjing* 210013, *China*)
* *Corresponding author, E-mail: xzp@usts.edu.cn*

Abstract: Addressing the issue of high computational complexity and limited extraction of spatio-temporal features in 3D convolutional neural networks for video super-resolution tasks, this paper introduced a novel lightweight video super-resolution reconstruction network based on hybrid spatio-temporal convolution. Firstly, a hybrid spatio-temporal convolution-based module was proposed to realize the enhancement of

收稿日期：2024-03-28；**修订日期**：2024-05-21.
基金项目：国家自然科学基金资助项目(No. 62002254)；江苏省自然科学基金资助项目(No. BK20200988)；苏州市科技计划项目(No. SNG-2023002)

the spatio-temporal feature extraction capability of the network as well as reduction of the computational complexity. Then, a similarity-based selective feature fusion module was proposed to further enhance the extraction capability of relevant features. Lastly, a motion compensation module based on the attention mechanism was designed to mitigate the effects of erroneous feature fusion to a certain extent. The experimental results demonstrate that the proposed network can achieve a favorable balance between video super-resolution performance and network complexity, and the 4-fold super-resolution reaches 8 frames per second on the benchmark dataset SPMCS-11. The proposed network meets the requirements for fast and accurate reasoning operations on edge devices.

Key words: video super-resolution; deep learning; 3D Convolutional Neural Network; feature fusion

1 引　言

超分辨率(Super-Resolution, SR)是图像处理领域中的一个基础性问题，其目标是将低分辨率(Low Resolution, LR)图像重建出对应的高分辨率(High Resolution, HR)图像。根据输入的帧数，超分辨率领域可以分为两类：单图像超分辨率(Single Image Super-Resolution, SISR)和视频超分辨率(Video Super-Resolution, VSR)。对于SISR来说，关键是利用图像的空间信息来修复缺失的细节。而对于VSR来说，如何充分挖掘相邻帧与目标帧之间的时间信息则是尤为重要的。如今，超分辨率技术应用广泛，主要应用领域包括医学影像处理[1]、视频监控[2]、遥感卫星图像处理[3]和超高清产业[4]等，因此具有重要的研究意义。

最近，深度学习在超分辨率领域取得了显著的进展，并成为研究的热点之一。DONG等人[5]提出了超分辨率重建卷积神经网络(Super-Resolution CNN, SRCNN)，该网络利用卷积层完成了从LR图像到HR图像的非线性映射。然而，由于SRCNN网络层数较少，无法充分提取到图像的特征，性能仍然有限。为了改进这个问题，Kim等人[6]提出更深层次的网络(Accurate Image Super-Resolution Using Very Deep Convolutional Networks, VDSR)，该网络采用了更多的卷积层，增加了感受野，并且通过残差训练方式，使得网络收敛速度更快。此外，周颖等人[7]将多尺度自适应注意力应用于图像的超分辨率重建，从而降低了图像的伪影，进一步丰富了图像细节。尽管这些网络取得了先进的性能，但它们在计算成本和内存占用方面较高，限制了它们在边缘设备上的应用。为了解决这个问题，一些轻量级网络被陆续提出，例如残差特征蒸馏网络(Residual Feature Distillation Network, RFDN)[8]，深度通道注意力网络(Deep Wise Channel Attention Network, DWCAN)[9]等。

视频超分辨率重建旨在利用输入的多个序列帧之间关联的时间、空间信息来重建图像。一般的方法常采用运动估计与运动补偿保持序列帧特征之间的对齐，以对齐后的序列帧作为输入，在超分网络中进行重建。这样，就可以在一定程度上减少由于运动引起的模糊和伪影。例如，Kappeler等人[10]提出了一种视频超分辨率重建网络(Video Super-Resolution Network, VSRNet)，该网络是图像超分算法SRCNN在视频上的扩展，增加了运动估计和运动补偿模块，实现了深度神经网络在视频超分辨率的首次应用。Huang等人[11]提出了一种双向循环卷积网络(Bidirectional Recurrent Convolutional Networks, BRCN)，可以对跨多帧的长期时间信息进行建模。王森等人[12]提出了一种多阶段帧对齐的视频超分辨率网络(Multi-Stage Frame Alignment Video Super-Resolution Network, MSVSR)，将VSR任务分解为去模糊、帧对齐以及特征重建三个阶段。此外，还有一类利用三维卷积神经网络(3D Convolutional Neural Networks, 3D CNN)提取视频序列帧间的时空特征完成重建的方法。在视频处理中，3D CNN相比于二维卷积神经网络(2D Convolutional Neural Networks, 2D CNN)拥有更多的优势。例如，Kim等人[13]受3D CNN固有的时空学习能力启发，提出了一种使用3D CNN的视频超分网络(Video Super-Resolution Network Using 3D Convolutional Neural

Networks, 3DSRNet)，该网络通过堆叠多个 3D 卷积层进行视频超分并避免了直接的运动对齐；Jo 等人[14]提出的动态上采样滤波视频超分辨率网络(Video Super-Resolution Network Using Dynamic Up-sampling Filter, DUF)构建了一个三维卷积模块，并采用动态上采样滤波器对目标帧进行重建，实现了良好的效果。然而，由于三维卷积内存占用过高且训练一个具有几十个三维卷积层的深层网络非常困难，目前大多数基于 3D CNN 的网络难以满足轻量级、高性能的需求。

为了解决上述问题，本文提出了一个新颖的混合时空卷积(Hybrid Spatial-Temporal Convolution, HSTC)来捕获深层的空间特征以及补偿多帧之间的时间信息。与以往的 3D CNN 不同，通过逐层堆叠 3D 卷积以同时处理空间和时间信号，所提出的 HSTC 将 3D CNN 与 2D CNN 进行联合，使得 3D CNN 主要捕获 2D CNN 中未捕获的时间信息，2D CNN 则用来提取静态但信息量很大的空间特征。与传统堆叠 3D 卷积的方式相比，HSTC 能够在很大程度上能够解决当前网络计算复杂度高以及特征提取能力有限的问题。同时，考虑到不合适的特征会影响特征融合的过程和最终的结果，本文进一步提出了一个基于相似性的选择性融合模块指导重建过程，能够强化相关特征并且排除不相关区域，进一步提高网络的性能。此外，本文还设计了一种基于注意力的运动补偿模块，进一步减轻错误特征融合的影响。实验结果表明，本文设计的轻量级网络在 VSR 任务中能够保持良好的性能并达到相当快的推理速度。

2　网络设计

2.1　网络结构

本文网络将 $N=2n+1$ 个 LR 帧的序列 $\{I_{t-n},\cdots,I_t,\cdots,I_{t+n}\}$ 作为输入，其中的 I_t 代表目标帧。视频超分辨率的目标是重建出 I_t 的高分辨率结果，由 I_{SR} 表示。如图 1 所示，整体架构由三个模块组成：运动补偿、时空特征提取、选择性特征融合。

(1) 运动补偿。在进入时空特征提取之前，每帧图像首先通过共享的 3×3 卷积层提取浅层特征 $\{F^i_{t-n},\cdots,F_t,\cdots,F^i_{t+n}\}$。然后，使用类似增强型可变形卷积的视频恢复网络(Video Restoration with Enhanced Deformable Convolutional

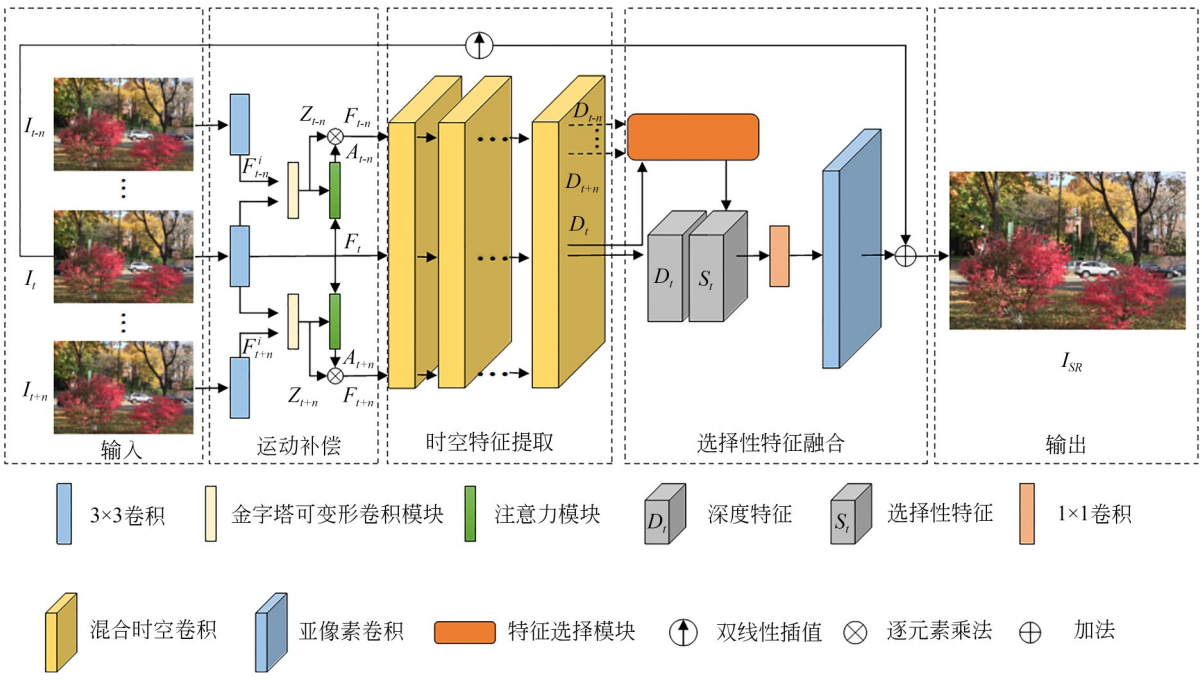


图 1　整体网络结构

Fig. 1　Overall network structure

Network，EDVR）的金字塔可变形卷积[15]，用于将每个相邻帧与目标帧 F_t 进行特征对齐，如图 2 所示。首先，金字塔可变形卷积在每个相邻帧和目标帧的浅层特征之间生成多尺度的特征。其次，第 l 级金字塔层的偏移和对齐特征分别使用来自第 $l-1$ 级的 2 倍上采样的偏移和对齐特征来进行预测。最后，每个相邻帧与目标帧都能够有效地对齐特征，从而实现由粗到细的运动补偿。整个运动补偿过程如式（1）～式（2）所示：

$$\Delta P_{t-n}^{l}=C_f([F_{t-n}^{i},F_t],(\Delta P_{t-n}^{l-1})^{\uparrow 2}), \tag{1}$$

$$Z_{t-n}=C_g(DConv(F_{t-n}^{i},\Delta P_{t-n}^{l}),((F_{t-n}^{i})^{l-1})^{\uparrow 2}), \tag{2}$$

式中：ΔP_{t-n}^{l} 表示第 l 级可变形卷积生成的偏移量；[，] 表示级联操作；$(\cdot)^{\uparrow 2}$ 表示上采样 2 倍；$(F_{t-n}^{i})^{l-1}$ 表示第 $t-n$ 帧图像在第 $l-1$ 级金字塔网络中提取的特征；C_f 和 C_g 表示由 3×3 卷积层组成的函数；$DConv$ 表示可变形卷积；Z_{t-n} 表示第 $t-n$ 帧对齐的特征。

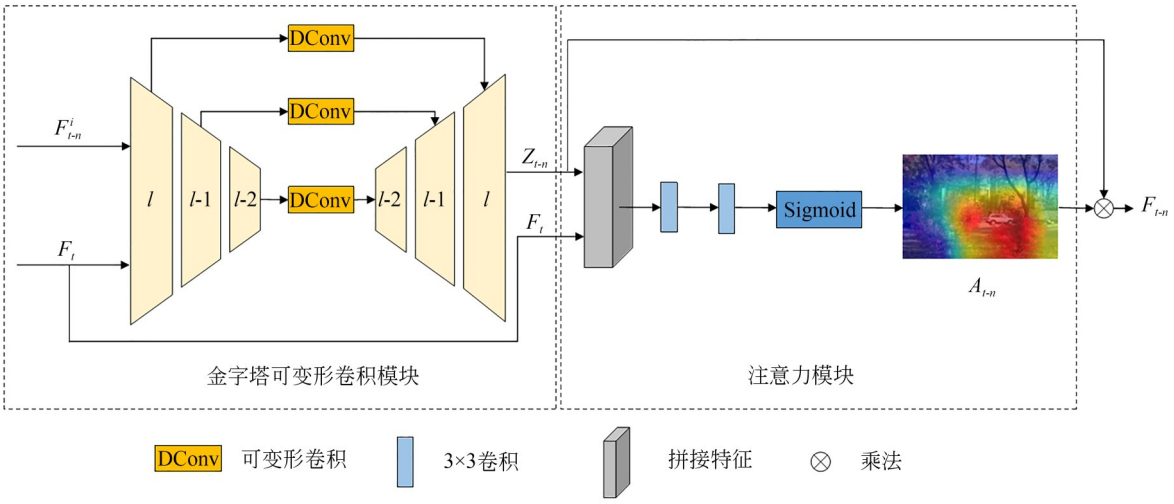


图 2　运动补偿结构

Fig. 2　Motion compensation structure

为了更好地利用每一帧的视觉信息以及避免错误特征进入合并过程，本文网络在运动补偿结构中引入了一种注意力机制，即双重注意力引导网络[16]。首先，给定对齐的特征 Z_{t-n}，与目标帧的浅层特征 F_t 合并在一起，然后经过两个 3×3 卷积和一个 sigmoid 函数激活后获得一个 128 维通道的注意力特征 A_{t-n}。最后，该对齐的特征和注意力特征进行逐元素相乘，得到融合运动信息的特征 F_{t-n}。该模块通过计算空间特征之间的相关性，可以有效地引导网络重建目标帧，提高网络的运动补偿能力。A_{t-n}，F_{t-n} 的表达式如式（3）～式（4）所示：

$$A_{t-n}=C_a([Z_{t-n},F_t]), \tag{3}$$

$$F_{t-n}=A_{t-n}\otimes Z_{t-n}, \tag{4}$$

式中：C_a 表示注意力运算函数；⊗ 表示逐元素乘法。

（2）时空特征提取模块。一方面，时间信息涉及视频中连续帧之间的信息和动态变化。例如，物体的运动轨迹、摄像机的移动、场景的变化等。另一方面，空间信息主要指的是每一帧内部的细节，如纹理、边缘、颜色等。该模块主要通过整合空间和时间信息来提取目标帧缺失的细节，从而有效地重建目标帧。其中，运动补偿后的特征表示为 $\{F_{t-n},\cdots,F_t,\cdots,F_{t+n}\}$，而经过轻量级的时空特征提取模块后的输出深度特征表示为 $\{D_{t-n},\cdots,D_t,\cdots,D_{t+n}\}$。

（3）选择性特征融合模块。如图 1 所示，首先，该模块将非目标帧的深度特征构建形成对应的特征字典，而目标帧的特征从字典中选择有用的特征并输出选择性特征 S_t。然后，将该特征与目标帧的深度特征 D_t 在通道维度上进行拼接，并将结果进行 1×1 的二维卷积运算，用于降低维度和提取相关信息。然后，在网络末端执行亚像素上采样操作以增加空间大小。最后，将预测图像

的残差连接到直接上采样的目标帧中，输出高分辨率图像 I_{SR}。

2.2　混合时空卷积

如图 3 所示，提出的 HSTC 以大小为 $T\times H\times W\times C$ 的多帧特征作为输入，并将共享权重的二维空间卷积分别应用于每一帧。其中，T 表示时间、H 表示高度、W 表示宽度、C 表示通道数。整个过程描述如式(5)所示：

$$F_t^{2D}=C_{SC}(F_t), \tag{5}$$

式中：t 表示时间索引；F_t 表示第 t 帧输入特征；C_{SC} 表示二维空间卷积运算函数；F_t^{2D} 表示第 t 帧输入特征经过二维空间卷积运算后的特征映射，该特征只包含每帧的空间信息。

为了对视频序列中的时间信息进行深度特征学习，网络引入了三维卷积运算层来学习连续帧中存在的时间相关性特征。如图 3 所示，取输入的三维特征为 $F=\{F_{t-n}^i,\cdots,F_t,\cdots,F_{t+n}^i\}$，通过沿着输入的时间维度进行卷积，输出特征图 F_t^{3D}，整个过程如式(6)所示：

$$F_t^{3D}=\kappa\otimes F, \tag{6}$$

式中：$\otimes$ 表示三维卷积运算；κ 表示三维卷积核。

然后，将二维空间卷积运算生成的特征映射与三维卷积生成的特征映射进行拼接。通过以上方式，所有通道的特征包含了两种信息：自身的空间信息、完全混合的时间信息。其次，使用一个二维卷积来进一步融合这些特征图的时空信息。相应地，融合后的特征映射与输入特征映射则具有相同的维度。简而言之，当以大小为 $T\times H\times W\times C$ 的多帧特征作为输入时，HSTC 将输出与输入相同大小的特征图。最后，HSTC 经过残差学习后的输出为 $F_t^*+F_t$。整个过程如式(7)所示：

$$F_t^*=C_{\text{fuse}}(\{F_t^{2D},F_t^{3D}\}), \tag{7}$$

式中：$\{F_t^{2D},F_t^{3D}\}$ 表示二维空间卷积和三维卷积计算后拼接的特征；C_{fuse} 表示二维卷积的融合函数；F_t^* 表示融合后的特征。

值得注意的是，由于跨不同帧的二维空间卷积的目标是提取其自身的深度空间特征，混合时空卷积采用参数共享的方案并不会降低网络的性能。同时，也不会导致内存复杂度过多地增加。因此，所提出的 HSTC 模块能够满足移动平台轻量级、高性能的需求。

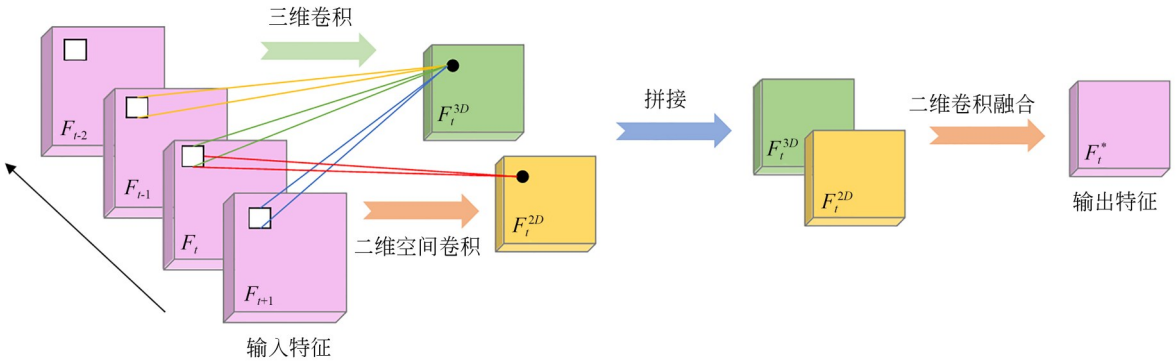


图 3　混合时空卷积

Fig. 3　Hybrid Spatial-Temporal Convolution

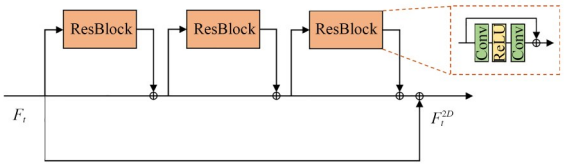


图 4　二维空间卷积

Fig. 4　2D Spatial Convolution

2.3　基于相似性的特征选择

基于相似性的特征选择模块由一个特征字典和一个基于相似性的寻址算子组成。其中，特征字典 L 用于表示来自 $\{D_{t-n},\cdots,D_t,\cdots,D_{t+n}\}$ 的特征，寻址算子则用于访问字典，具有与文献[17]中类似的结构和操作。

(1) 基于字典的选择性表示。特征字典被设计为矩阵 $L\in R^{M\times C}$，表示为 M 个非目标帧的特征向量且每个向量的通道维度为 C。同时，输入特征 D_t 每个像素上的特征向量 q 具有与矩阵 L 中的每个特征向量相同的维度，即 $q\in R^C$。若设行

向量 $l_i \forall [M]$，则 l_i 表示矩阵 L 第 i 行的特征字典项，$[M]$ 表示从 1 到 M 的整数集合。对于给定查询 q，当访问特征字典时，可以根据它与每个特征字典项 l_i 之间的相似性来计算权重 $\hat{w} \in R^{1\times M}$，那么将输出对应的选择性特征向量 $\hat{q}$。整个过程描述如式(8)所示：

$$\hat{q} = \hat{w}L = \sum_{i=1}^{M} \widehat{w}_i l_i, \tag{8}$$

式中：$\hat{w}$ 表示具有总和为 1 的非负值行向量，$\widehat{w}_i$ 是 $\hat{w}$ 中第 i 个值，w 表示全为 1 的初始权重。

(2) 基于相似性的寻址操作符。本文设计了一个字典寻址的子模块，能够从内容可寻址特征字典中找到与输入查询相关的信息。该模块基于字典项和查询 q 的相似性来计算相似性权重 $\hat{w}$。如图 5 所示，该模块通过 *softmax* 运算计算每个权重项 $\widehat{w}_i$。整个过程描述如式(9)～式(10)所示：

$$\hat{w} = \frac{\exp(d(q, l_i))}{\sum_{j=1}^{M} \exp(d(q, l_i))}, \tag{9}$$

$$d(q, l_i) = \theta(q)^{\mathrm{T}} \phi(l_i), \tag{10}$$

式中：$d(q, l_i)$ 表示计算 q 与所有字典项之间的相似性的函数；$d(\cdot, \cdot)$ 表示嵌入的高斯函数；$\theta(q) = W_\theta q$；$\phi(l_i) = W_\phi l_i$；W_θ 和 W_ϕ 表示 1×1 卷积的权值；exp(·)表示指数函数；T 表示转置操作。

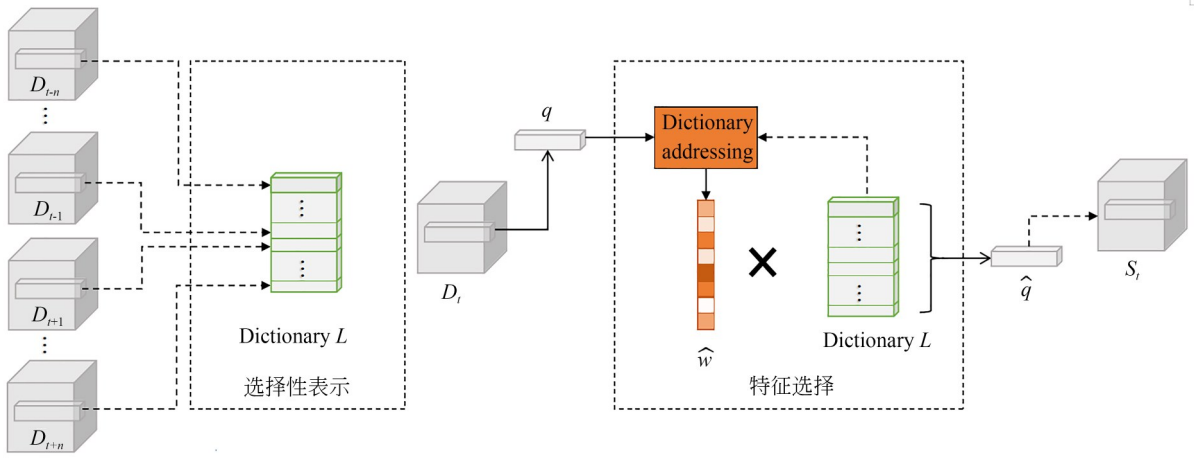


图 5　基于相似性的特征选择

Fig. 5　Similarity-based feature selection

根据以上等式，基于相似性的特征选择模块能够选择与查询最相似的字典项，用来获得非目标中帧所需的细节表示。一方面，所选择的相似性表示形成特征图 S_t，能够补偿和重建目标，同时避免不相关的混淆特征阻碍细节合并的结果。另一方面，所提出的基于相似性的特征选择模块采用非局部操作，依靠基于字典的操作，可以应用于更一般的尺度和更灵活的设置。

2.4　损失函数

本文网络采用 Charbonnier 损失作为图像重建损失去衡量 SR 图像与真实高分辨率 HR 图像之间的差异。在训练过程中，通过反向传播算法不断调整网络的参数，使得 Charbonnier 损失最小化，从而提高超分辨率图像的重建质量。其中，Charbonnier 损失的表达式如式(11)所示：

$$Loss_{Lc} = \frac{1}{N} \sum_{i=1}^{N} \sqrt{\left(I_{HR}^{i} - I_{SR}^{i}\right)^2 + \varepsilon^2}, \tag{11}$$

式中：I_{SR}^{i} 和 I_{HR}^{i} 表示生成的第 i 张 SR 图片和相对应的第 i 张 HR 图片；N 为样本数，ε 是一个很小的常数，设置为 10^{-3}。

3　实验结果和分析

3.1　实验数据与评价指标

本文使用 Vimeo-90K 数据集[18]作为网络的训练集。该数据集是一个大规模、高质量的视频数据集，包含 64 612 个 7 帧的视频序列，每帧的像素大小为 448×256。通过对高分辨率的视频帧使用双三次插值下采样，得到对应的低分

辨率视频帧。因此，在4倍超分辨率上，使用的输入帧是112×64 pixels的低分辨率视频帧。

测试数据集包括Vimeo-90K-T[18]，Vid4[19]，SPMCS[20]。其中，Vid4数据集包含4个不同场景的视频片段，具有多种运动和遮挡，但视频帧之间的运动幅度较小，平均流大小（像素/帧）[21]较低，并且在高分辨率视频帧的部分区域上存在一些伪影。SPMCS数据集由30个视频序列组成，每个视频序列包含分辨率为540×960 pixels的31个连续帧。但在实际测试时，本文和许多方法一样选择由其中的11张组成的SPMCS-11数据集，这是为了选择更干净的高分辨率视频帧，减少对测试结果的干扰。此外，本文还使用了Vimeo-90K-T数据集作为测试数据集。该数据集包含多种多样的对象与运动环境，包括慢速动作、中速动作和快速动作三种不同平均流大小（像素/帧）的序列。而且，该数据集还包含丰富而复杂的光线变化，为算法带来巨大的挑战。通过这些数据集，可以满足区分多种算法优劣的需求。因此，通过使用这些数据集进行训练和测试，能够评估和比较不同算法的性能和优劣。

评价方面，在由RGB空间变换到YCbCr空间的Y单通道上分别评估峰值信噪比（Peak Signal to Noise Ratio，PSNR）[22]和结构相似性（Structural Similarity，SSIM）[23]等指标。

3.2 实验设置

本文网络结构如表1所示。其中，$C_{TC}(\cdot)$表示时间×高度×宽度×卷积核数量，其余则表示高度×宽度×卷积核数量。在训练过程中，一组相关的7张连续低分辨率的视频帧传入，并从视频帧中裁剪64×64像素大小的图片输入到网络中。批次大小则设置为16，而时空特征提取模块中总共使用了10个HSTC，中间的特征图通道则设置为128。同时，网络采用Adam优化器[24]，一阶矩和二阶矩指数衰减率分别设置为$\beta_1=0.9$，$\beta_2=0.999$，数值稳定常数$\varepsilon=10^{-8}$。学习率初始设置为10^{-4}，然后在每隔75个训练周期后变为原来的一半。本文算法模型在NVIDIA RTX 3090Ti GPU上进行训练和测试，在Pytorch1.10.0框架上部署实施。

表1 网络结构

Tab. 1 Architecture of network

模块	函数名	卷积核大小
运动补偿	$C_f(\cdot)$	3×3×128
	$C_g(\cdot)$	3×3×128
	DConv	3×3×128
时空特征提取	$C_a(\cdot)$	3×3×128
	$C_{SC}(\cdot)$	3×3×128
	$C_{TC}(\cdot)$	3×3×3×128
	$C_{fuse}(\cdot)$	3×3×128
选择性特征融合	$\theta(\cdot)$	3×3×128
	$\phi(\cdot)$	1×1×128
	$C_e(\cdot)$	1×1×128
	Up sampling	3×3×48

3.3 消融实验

（1）模块有效性分析。本文研究了设计的网络架构，并验证了不同模块在整个模型中的重要性，包括三维卷积、二维空间卷积、HSTC和选择性特征融合策略。在这些比较中，所有的模型都进行了重新训练，并在SPMCS-11数据集中进行了测试。为了进行公平的比较，除了Deep-TC-VSR之外，其他模型通过添加更多的层数，与最终的模型S-HSTC-VSR具有相似的参数数量。

以下是消融研究中各种模型的介绍：（1）S-HSTC-VSR表示所提方法的完整模型；（2）HSTC-VSR表示去除S-HSTC-VSR中的基于相似性的特征选择，将输出的特征映射$\{D_{t-n},\cdots,D_t,\cdots,D_{t+n}\}$直接合并，最后输入到上采样重建网络；（3）S-TC-VSR表示在所提出的HTSC中去除二维空间卷积；（4）S-SC-VSR表示在所提出的HTSC中去除三维卷积；（5）TC-VSR表示一个基于多个三维卷积的融合网络，即去除原有的选择性特征融合模块和二维空间卷积；（6）Deep-TC-VSR表示使用了相比于TC-VSR更多的三维卷积的网络。

表2展示了不同模型在数据集SPMCS-11上的PSNR和SSIM的对比，并对每个测试数据集上PSNR和SSIM性能指标表现最好的算法进行加粗处理，对次优算法进行下划线标注。一方面，可以看出HSTC模块是一种针对时空特征提取极为有效的机制，即使Deep-TC-VSR具有更多的三维卷积和更多的参数，与TC-VSR相比，

表 2　不同模型在数据集 SPMCS-11 上的定量比较

Tab. 2　Quantitative comparison of different activation functions on the SPMCS-11 dataset

模型	三维卷积	二维空间卷积	特征选择模块	PSNR	SSIM
TC-VSR	√			29.47	0.869 9
Deep-TC-VSR	√			29.79	0.874 8
S-TC-VSR	√		√	29.72	0.873 4
S-SC-VSR		√	√	29.61	0.871 3
HTSC-VSR	√	√		29.59	0.873 5
S-HTSC-VSR(Ours)	√	√	√	**30.51**	**0.880 9**

提升的性能依然很小。与 TC-VSR 相比，HSTC-VSR 由于结合了二维空间和三维时间信息，获得了更好的结果。同样，S-HSTC-VSR 也获得了比 S-TCVSR 和 S-SC-VSR 更好的性能。其次，去除二维空间卷积后，HTSC-VSR 的性能比 Deep-TC-VSR 的性能要差，这很大程度上是因为 Deep-TC-VSR 拥有更多的三维卷积。最后，为了进一步验证结果，本文进一步展现了它们之间的视觉对比，其结果如图 6 所示。虽然与 TC-VSR 相比，Deep-TC-VSR 使边缘更加清晰，但仍然存在一些混乱的纹理。仅具有空间分量的网络（S-SC-VSR）的恢复视频利用来自同一帧中相邻区域的内部细节来恢复更好的静止纹理。仅具有时间分量的网络（S-TC-VSR）在恢复具有边缘和遮挡的区域方面则更为强劲。实验结果表明，所提出的网络（S-HSTC-VSR）能够与空间和时间信息相结合，显示出更为清晰的纹理。

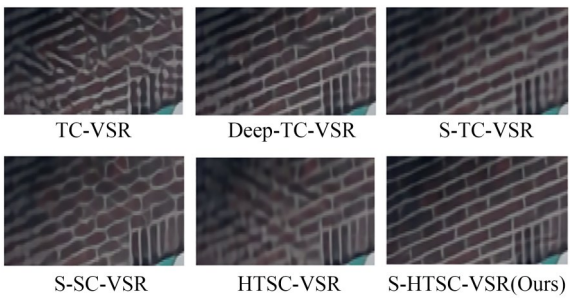


图 6　本文网络及其变体的可视化结果

Fig. 6　Visual results of our network and its variants

另一方面，为了证明选择性特征融合策略的效果，本文比较了 S-HSTC-VSR 和 HSTC-VSR。结果显示，采用选择性特征融合策略的模型可以获得比直接融合策略更高的性能指标。与三维卷积融合策略相比，S-TC-VSR 的性能也同样优于 TC-VSR。从图 6 中可以发现，采用选择性特征融合策略的模型能够有效地缓解了图像中纹理的混淆。实验结果表明，所提出的选择性特征融合策略有助于避免特征混淆和抑制帧间不相关的特征。

（2）参数设置研究分析。该部分主要探讨不同深度和宽度对网络性能的影响，如表 3 所示。深度表示混合时空卷积模块的数量，宽度表示卷积核的通道数量。其中，网络是在 SPMCS-11 数据集进行的 4 倍超分辨率。为了在网络性能和计算复杂度之间取得一定的平衡，本文尝试网络在深度为 8 和 10，宽度在 64 和 128 进行相应的实验。实验结果表明，网络深度为 10、宽度为 128 时，模型参数量和性能综合最好。因此，最终网络深度和宽度分别设置为 10 和 128。

表 3　不同宽度和深度的网络性能

Tab. 3　Network performance of different widths and depths

深度	宽度	参数量/M	PSNR/dB	SSIM
8	64	3.9	30.21	0.870 1
8	128	5.2	30.27	0.874 4
10	64	6.8	30.43	0.875 2
10	128	9.7	**30.51**	**0.880 9**

（3）损失函数有效性分析。本文网络使用不同的损失函数进行训练的定量比较如表 4 所示，最优的指标进行了加粗处理。实验表明，与 MSE 损失相比，L1 损失能够保留更多的细节。此外，用 Charbonnier 损失训练的网络获得了更好的性能。因此，网络采用 Charbonnier 损失来训练模型。

表4　不同损失函数在Vid4数据集上所有视频帧的平均值

Tab. 4　Average value of all video frames of different Loss Functions on the Vid4 dataset

	MSE loss	L1 loss	Charbonnier loss
PSNR	27.28	27.36	**27.43**

3.4　对比实验

(1) 网络重建性能对比。在本节中,网络在Vid4、SPMCS-11和Vimeo-90K-T等3个数据集上进行了实验,并与多个算法进行了实验,其中包括Bicubic、残差通道注意超分网络(Image Super-Resolution Using Very Deep Residual Channel Attention Networks, RCAN)[25],DUF[14]、时空可变形对齐视频超分网络(Temporally Deformable Alignment Network for Video Super-Resolution, TDAN)[21]、视频超分Transformer(VSR-Transformer)[26]以及BasicVSR++[27]。表5~表7展示了不同方法在Vid4数据集、SPMCS-11数据集和Vimeo-90K-T数据集上进行4倍超分的定量对比结果。标粗的指标表示最优的结果,标下划线的指标表示次优的结果。由表5可以观察到,所提网络在"Calendar"和"Walk"的视频序列中实现了最佳的SSIM值。同时,在其他序列上,所提网络优于大部分的超分辨率方法而仅次于DUF,这可能是由于原始的高分辨率图像中部分区域存在的伪影导致指标值不高。而在"City"与"Foliage"视频帧序列中数值稍逊于DUF,只取得了次优的数值,其主要原因在于DUF算法主要是使用Vid4数据集进行训练的,对Vid4数据集的测试会更具优势;由表6可以观察到,对比网络BasicVSR++,DUF,TDAN等视频超分方法,所提网络能够学习更为精确的高频特征,在PSNR方面平均分别领先0.14 dB,1.09 dB,1.53 dB,在SSIM方面平

表5　×4尺度下不同算法在Vid4数据集上的定量对比(PSNR(dB)/SSIM)

Tab. 5　Quantitative comparisons of different algorithms for scale factor ×4 on Vid4 dataset(PSNR(dB)/SSIM)

片段名	Bicubic	RCAN[25]	DUF[14]	TDAN[21]	VSR-Transformer[26]	BasicVSR++[27]	Ours
Calendar	20.39/0.572 0	22.31/0.724 8	24.04/0.811 0	23.20/0.768 9	24.14/0.815 7	**24.23**/0.820 9	24.20/**0.821 2**
City	25.16/0.602 8	26.07/0.693 8	**28.27/0.831 3**	27.18/0.771 6	27.87/0.811 4	28.01/0.813 7	28.03/0.814 1
Foliage	23.47/0.566 6	24.69/0.662 8	**26.41/0.770 9**	25.64/0.728 4	26.29/0.761 3	26.34/0.765 4	26.39/0.766 5
Walk	26.10/0.797 4	28.64/0.871 8	30.30/0.914 1	29.80/0.894 0	30.91/0.910 9	**31.11**/0.915 4	31.09/**0.915 7**
Average	23.78/0.634 7	25.43/0.738 3	27.26/**0.831 8**	26.46/0.790 7	27.30/0.824 8	27.42/0.828 9	**27.43**/0.829 4

表6　×4尺度下不同算法在SPMCS-11数据集上的定量对比(PSNR(dB)/SSIM)

Tab. 6　Quantitative comparisons of different algorithms for scale factor ×4 on SPMCS-11 dataset(PSNR(dB)/SSIM)

片段名	Bicubic	RCAN[25]	DUF[14]	TDAN[21]	VSR-Transformer[26]	BasicVSR++[27]	Ours
Car_05	27.75/0.782 5	29.84/0.848 3	30.77/0.870 5	30.59/0.865 3	32.13/0.903 2	32.31/0.905 4	**32.42/0.906 3**
hdclub_003	19.42/0.486 3	20.39/0.610 0	22.06/0.742 9	21.34/0.687 9	22.11/0.738 7	**22.19/0.744 3**	22.17/0.741 9
hitachi_isee5	19.61/0.593 8	23.58/0.837 1	25.75/0.892 7	24.59/0.856 7	26.50/0.906 9	26.73/0.909 7	**26.74/0.912 3**
hk004_001	28.54/0.800 3	31.72/0.862 8	32.96/0.898 4	32.27/0.882 5	33.48/0.904 6	33.59/**0.905 1**	**33.66**/0.904 5
HKVTG_004	27.46/0.683 1	28.77/0.765 0	29.15/0.785 5	29.11/0.778 8	29.57/0.798 3	**29.60**/0.798 7	29.55/**0.801 3**
jvc_009	25.40/0.755 8	28.29/0.872 2	29.17/0.895 9	28.90/0.883 2	30.46/0.919 5	30.74/0.921 1	**30.91/0.921 6**
NYVTG_006	28.45/0.801 4	30.99/0.886 0	32.32/0.905 8	31.90/0.899 6	33.32/0.925 1	33.56/0.926 9	**34.11/0.927 4**
PRVTG_012	25.63/0.713 6	26.63/0.781 1	27.35/0.816 4	27.16/0.805 6	27.67/0.825 3	27.79/**0.828 1**	**27.84**/0.827 4
RMVTG_011	23.96/0.657 3	26.05/0.757 4	27.53/0.811 5	26.95/0.792 4	27.71/0.819 7	27.81/0.823 4	**27.94/0.825 2**
veni3_011	29.47/0.897 9	34.54/0.962 5	34.64/0.967 6	34.68/0.964 5	36.53/0.974 5	36.57/0.974 8	**37.16/0.975 2**
veni5_015	27.41/0.848 3	31.01/0.926 2	31.89/0.936 7	31.30/0.927 5	32.77/0.944 9	**33.17/0.947 3**	33.12/0.946 6
Average	25.73/0.739 1	28.35/0.828 1	29.42/0.865 9	28.98/0.849 5	30.20/0.878 2	30.37/0.880 4	**30.51/0.880 9**

表 7　×4 尺度下不同算法在 Vimeo-90K-T 数据集上的定量对比（PSNR（dB）/SSIM）

Tab. 7　Quantitative comparisons of different algorithms for scale factor ×4 on Vimeo-90K-T dataset（PSNR（dB）/SSIM）

算法	慢速运动	中速运动	快速运动	Average
Bicubic	29.34/0.833 0	31.29/0.870 8	34.07/0.905 0	31.32/0.868 4
RCAN[25]	32.92/0.902 8	35.33/0.926 5	38.45/0.945 3	35.32/0.924 5
DUF[14]	33.38/0.910 7	36.69/0.944 2	38.86/0.950 8	36.35/0.938 3
TDAN[21]	33.17/0.906 5	36.05/0.936 9	38.70/0.949 1	35.87/0.932 5
VSR-Transformer[26]	34.43/0.923 2	37.69/0.951 7	40.26/0.961 3	37.42/0.947 3
BasicVSR＋＋[27]	**34.58/0.925 6**	37.75/0.952 7	40.49/0.962 4	37.52/0.948 6
Ours	34.53/0.924 6	**37.81/0.953 5**	**40.56/0.963 3**	**37.56/0.949 0**
片段数量	1 616	4 983	1 225	7 824
平均流大小	0.6	2.5	8.3	3.0

均分别领先 0.000 5，0.015 0，0.031 4 dB；Vimeo-90K-T 数据集的定量对比结果如表 7 所示。在现实世界中拍摄的视频中，当物体运动幅度更大且速度更快时，视频序列中包含的时间信息也会相应增加。然而，这也意味着视频超级分辨率网络在重构过程中需要处理更多的时间信息，从而使其重构难度更大。从表 7 中观察到，所设计的网络在中速运动、快速运动两种运动幅度上取得了最高的 PSNR 和 SSIM 结果，在慢速运动上得到了次优结果。因为快速运动的组中平均流大小大幅度超过了之前测试的两个数据集，所以组内序列具有更大的运动幅度，变化也更为多样。因此，这些数据足以证明所提网络能够更有效地提取相邻帧的运动特征以及帮助目标帧进行重建恢复。

（2）网络重建视觉对比。为了观察视频超分辨率的重建结果，本文选取先进算法以及双三次插值法在 4 倍超分辨率上，进行部分视频的重建视觉对比实验，包括 Vid4 数据集中的“Calendar”、SPMCS-11 数据集中的“jvc_009”以及 Vimeo-90K-T 数据集中的“0096_0725”。主观视觉结果的对比如图 7 所示（彩图见期刊电子版），红色方框处标记出不同方法间具有明显差异的位置，GT（Ground Truth）代表原始的高分辨率影像。在“Calendar”视频序列中，仅有所提网络在没有变形的情况下恢复了日历中的日期，而其他

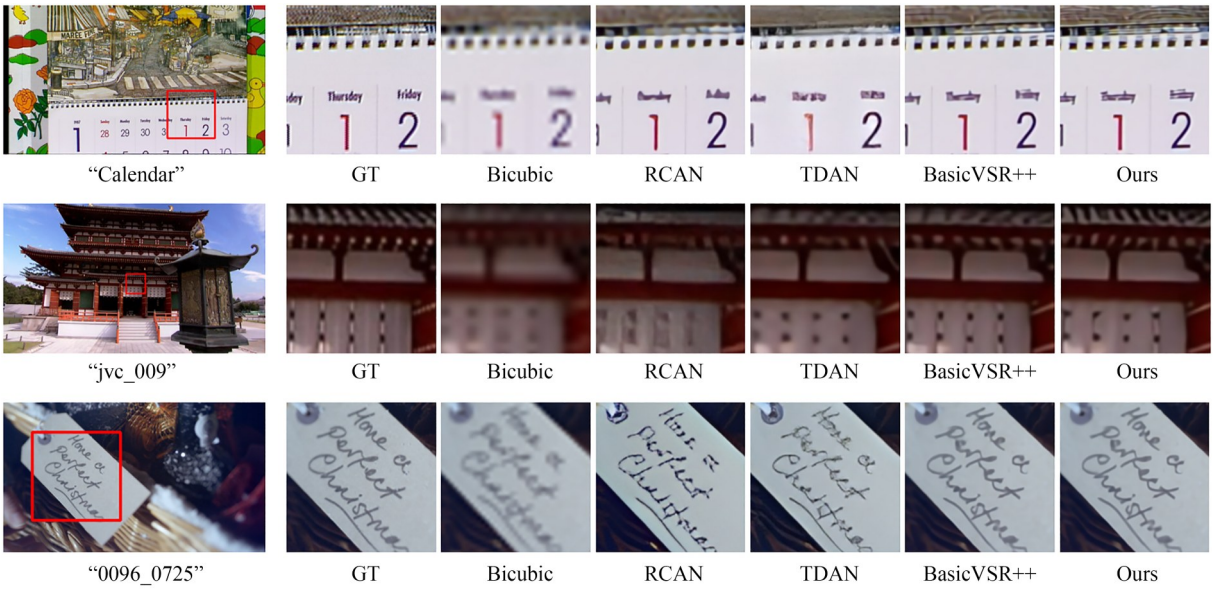


图 7　在 3 个数据集（×4）上先进算法与所提网络的重建视觉对比

Fig. 7　Reconstruction visual comparisons of the state-of-the-art algorithms and proposed network on three datasets for ×4 SR

方法则出现了一定范围的模糊区域或不希望产生的伪影。此外，所提网络在“jvc_009”视频序列中显示了更为精确的建筑图像，而其他方法恢复的条纹方向较为不正确。而对于图中第三行的文字，所以网络重建的高分辨率图像更加贴近真实的视频帧，能够恢复出正确的文字边缘。因此，所提网络能够从低分辨率视频帧的特征中更有效地利用全局视频帧的特征，重建出纹理细节更为出色的图像。

（3）网络泛化性对比。目前，绝大部分的算法都是基于模拟的数据集进行训练，而现实世界中的数据集往往会被多种因素影响，例如相机的抖动、其他噪声等，因此常规的视频超分模型对恢复真实的视频存在很大的难度。为了进一步证明所提网络的有效性，本文在Liao等人发表的真实数据集[28]的基础上进行了对比实验，以评估网络的泛化能力。由于没有给出真实视频帧的标准参考HR，因此本文均通过两种无参考质量评估指标（Natural Image Quality Evaluator，NIQE）[29]，（Spatial Spectral Entropy-based Quality，SSEQ）[30]来评估各种算法的表现。从表8可以看到，所提网络在NIQE和SSEQ指标下表现出了最佳的性能，能够证明该网络模型具有强大的泛化能力，并在真实数据集上取得良好的表现。

表8　在真实世界数据集上的定量对比

Tab. 8　Quantitative comparisons on the real-world dataset

评估指标	Bicubic	RCAN[25]	TDAN[21]	BasicVSR++[27]	Ours
NIQE↓	7.58	6.29	6.56	<u>6.11</u>	**6.05**
SSEQ↓	54.40	46.32	44.26	<u>41.17</u>	**40.59**

（4）网络运行时间对比。网络运行时间对于轻量级视频超分算法也是极为重要的指标。为了进一步验证本文所提网络是轻量级的网络，选取主流网络在SPMCS-11数据集（×4）上进行重建速度的对比实验，测试视频帧的尺寸为240×134像素，并去除掉第一个视频序列的推理时间，因为第一个视频序列的推理时间可能会包括网络加载时间。重复运行10次后得出测试的平均推理时间，如表9所示。可以看出，所提网络的重建速度次优，参数量和模型计算量分别仅为9.7 M，19.04×10^9，远远优于传统基于普通3D卷积的3DSRNet和当前流行的VSR-Transformer。而且，在平均运行时间方面分别只有3DSRNet、VSR-Transformer的15%，10%。同时，在PSNR和SSIM指标方面，比先进算法BasicVSR++分别高出0.14 dB，0.000 5。因此，能够证明所提网络是轻量级的网络且具有显著优势。

表9　在SPMCS-11数据集（×4）上的平均运行时间

Tab. 9　Average running time on SPMCS-11 dataset for ×4 SR

算法	PSNR/dB	SSIM	参数量/M	FLOPs/10^9	平均运行时间/s
RCAN[25]	28.35	0.828 1	15.6	261.46	1.586
DUF[14]	29.42	0.865 9	**5.8**	92.97	0.573
3DSRNet[13]	28.98	0.849 5	15.9	127.49	0.778
VSR-Transformer[26]	30.20	0.878 2	43.8	834.01	1.153
BasicVSR++[27]	<u>30.37</u>	<u>0.880 4</u>	<u>6.4</u>	**11.07**	**0.067**
Ours	**30.51**	**0.880 9**	9.7	<u>19.04</u>	<u>0.115</u>

4　结　论

针对三维卷积神经网络在视频超分辨率任务上具有较高的计算复杂度以及提取的时空特征有限的问题，本文设计了一种基于混合时空卷积的轻量级视频超分辨率重建网络。首先，提出了一个基于混合时空卷积的模块，实现了网络时空特征提取能力的提升以及计算复杂度的降低；然后，提出了一个基于相似性的选择性特征融合模块，进一步增强了相关特征的提取能力；最后，

设计了一种基于注意力机制的运动补偿模块,在一定程度上减轻了错误的特征融合的影响。实验结果表明,所提网络可以在视频超分辨率性能和网络复杂度之间取得很好的平衡,而且在基准数据集SPMCS-11上4倍超分辨率达到8 frame/s,能够满足实际的应用需求。

参考文献:

[1] 林毅,周芃,陈彦明. 基于语义注意力的医学图像超分辨率方法[J]. 计算机科学, 2023, 50(S2): 1017-1022.
LIN Y, ZHOU P, CHEN Y M. Medical image super-resolution method based on semantic attention [J]. *Computer Science*, 2023, 50 (S2): 1017-1022. (in Chinese)

[2] KHAN M A, JAVED K, KHAN SALI, *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.

[3] 易见兵, 陈俊宽, 曹锋, 等. 轻量级重参数化的遥感图像超分辨率重建网络设计[J]. 光学 精密工程, 2024, 32(2): 268-285.
YI J B, CHEN J K, CAO F, *et al*. Design of lightweight re-parameterized remote sensing image super-resolution network [J]. *Opt. Precision Eng.*, 2024, 32(2): 268-285. (in Chinese)

[4] 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.

[5] DONG C, LOY C C, HE K, *et al*. Learning a Deep Convolutional Network for Image Super-Resolution[C]. *Proceedings of the* 2014 *European Conference on Computer Vision. Cham*: Springer, 2014: 184-199.

[6] KIM J, LEE J K, LEE K M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks[C]. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). *Las Vegas*, *NV*, *USA*. IEEE, 2016: 1646-1654.

[7] 周颖, 裴盛虎, 陈海永, 等. 基于多尺度自适应注意力的图像超分辨率网络[J]. 光学 精密工程, 2024, 32(6): 843-856.
ZHOU Y, PEI S H, CHEN H Y, *et al*. Image super-resolution network based on multi-scale adaptive attention[J]. *Opt. Precision Eng.*, 2024, 32(6): 843-856. (in Chinese)

[8] LIU J, TANG J, WU G S. Residual Feature Distillation Network for Lightweight Image Super-Resolution[C]. *BARTOLI A*, *FUSIELLO A. European Conference on Computer Vision. Cham*: Springer, 2020: 41-55.

[9] YASIR M, ULLAH I, CHOI C. Depthwise channel attention network (DWCAN): an efficient and lightweight model for single image super-resolution and metaverse gaming[J]. *Expert Systems*, 2024, 41(4): e13516.

[10] 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.

[11] HUANG Y, WANG W, WANG L. Bidirectional recurrent convolutional networks for multi-frame super-resolution[J]. *Advances in Neural Information Processing Systems*, 2015, 28.

[12] 王森, 祝阳, 张印辉, 等. 多阶段帧对齐的视频超分辨率重建网络[J]. 光学 精密工程, 2023, 31 (16): 2430-2443.
WANG S, ZHU Y, ZHANG Y H, *et al*. Multi-stage frame alignment video super-resolution network[J]. *Opt. Precision Eng.*, 2023, 31(16): 2430-2443. (in Chinese)

[13] KIM S Y, LIM J, NA T, *et al*. 3dsrnet: Video super-resolution using 3d convolutional neural networks [EB/OL]. [2019-01-20]. https://arxiv.org/pdf/1812.09079.pdf.

[14] JO Y, OH S W, KANG J, *et al*. Deep Video Super-Resolution Network Using Dynamic Upsampling Filters Without Explicit Motion Compensation[C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City*, *UT*, *USA*. IEEE, 2018: 3224-3232.

[15] WANG X T, CHAN K C K, YU K, *et al*. EDVR: Video Restoration with Enhanced Deformable Convolutional Networks [C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*). *Long Beach*, *CA*, *USA*. IEEE, 2019: 1954-1963.

[16] YAN Q S, GONG D, SHI J Q, *et al*. Dual-attention-guided network for ghost-free high dynamic

range imaging[J]. *International Journal of Computer Vision*, 2022, 130(1): 76-94.

[17] GONG D, LIU L Q, LE V, *et al*. Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder For Unsupervised Anomaly Detection [C]. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *Seoul*, *Korea* (*South*). IEEE, 2019: 1705-1714.

[18] 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.

[19] 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.

[20] TAO X, GAO H Y, LIAO R J, *et al*. Detail-revealing deep video super-resolution [C]. 2017 *IEEE International Conference on Computer Vision* (*ICCV*). *Venice*, *Italy*. IEEE, 2017: 4482-4490.

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

[22] GAO X B, LU W, TAO D C, *et al*. Image quality assessment based on multiscale geometric analysis[J]. *IEEE Transactions on Image Processing*: *a Publication of the IEEE Signal Processing Society*, 2009, 18(7): 1409-1423.

[23] WANG Z, BOVIK A C, SHEIKH H R, *et al*. Image quality assessment: from error visibility to structural similarity[J]. *IEEE Transactions on Image Processing*, 2004, 13(4): 600-612.

[24] KINGMA D P, BA J. Adam: A method for stochastic optimization [EB/OL]. [2017-01-30]. https://arxiv.53yu.com/pdf/1412.6980.pdf.

[25] ZHANG Y L, LI K P, LI K, *et al*. Image Super-Resolution Using Very Deep Residual Channel Attention Networks [C]. *European Conference on Computer Vision*. *Cham*: *Springer*, 2018: 294-310.

[26] CAO J Z, LI Y W, ZHANG K, *et al*. Video super resolution transformer [EB/OL]. [2022-02-08]. https://arxiv.org/pdf/2106.06847.pdf.

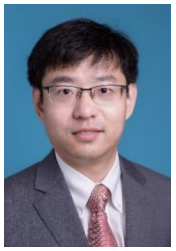
[27] CHAN K C K, ZHOU S C, XU X Y, *et al*. BasicVSR: improving video super-resolution with enhanced propagation and alignment [C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *New Orleans*, *LA*, *USA*. IEEE, 2022: 5962-5971.

[28] LIAO R J, TAO X, LI R Y, *et al*. Video super-resolution via deep draft-ensemble learning [C]. 2015 *IEEE International Conference on Computer Vision* (*ICCV*). *Santiago*, *Chile*. IEEE, 2015: 531-539.

[29] WANG Z, BOVIK A C, SHEIKH H R, *et al*. Image quality assessment: from error visibility to structural similarity[J]. *IEEE Transactions on Image Processing*, 2004, 13(4): 600-612.

[30] LIU L X, LIU B, HUANG H, *et al*. No-reference image quality assessment based on spatial and spectral entropies [J]. *Signal Processing*: *Image Communication*, 2014, 29(8): 856-863.

作者简介：



夏振平(1985—),男,江苏兴化人,博士,副教授,硕士生导师,2014年于东南大学获得博士学位,主要从事视觉相关的显示图像质量测量、评价和优化方面的研究。E-mail: xzp@usts.edu.cn

陈　豪(2000—),男,四川广安人,硕士研究生,2021年于苏州科技大学获得学士学位,主要从事深度学习、数字图像处理方面的研究。E-mail: cvchen1595@163.com。