文章编号　1004-924X(2023)16-2430-14

多阶段帧对齐的视频超分辨率重建网络

王　森，祝　阳，张印辉*，王庆健，何自芬
(昆明理工大学 机电工程学院，云南 昆明 650500)

摘要：视频超分辨率(Video-Super Resolution，VSR)旨在将低分辨率视频帧序列重建为高分辨率视频帧序列。相较于图像超分辨率，VSR由于增加了时间维度的信息，因此通常需要依赖邻近帧高度相关信息实现当前帧的重建。如何对齐相邻帧，并获取帧间高度相关信息，是VSR任务关注的重点问题。本文将VSR任务分为去模糊、对齐、重建三个阶段。在去模糊阶段，将当前帧与相邻帧进行预对齐，获取与当前帧高度相关的特征信息，通过强化当前帧的细节以便实现初始阶段更多特征信息的提取。在对齐阶段，通过对输入特征进行二次对齐操作，利用相邻帧中高度相关信息进一步强化当前帧中特征信息。在重建阶段，通过聚合原始低分辨率帧以在网络末端提供更多特征信息。本文利用多层感知机(Multi-Layer Perceptron，MLP)代替传统卷积操作构造特征提取模块，同时对生成的特征信息进行二次对齐，以细化图像特征获得更优的视频帧重建效果。实验结果表明，本文提出的算法在多种公开数据集上的视频帧序列重建精度更高的同时，也取得了更少的网络参数量和更连贯的视频序列重建表现。

关　键　词：计算机视觉；视频超分辨率；多层感知机；注意力机制；光流；帧对齐
中图分类号：TP391　　**文献标识码**：A　　**doi**：10.37188/OPE.20233116.2430

Multi-stage frame alignment video super- resolution network

WANG Sen，ZHU Yang，ZHANG Yinhui*，WANG Qingjian，HE Zifen

(*Faculty of Mechanical and Electrical Engineering*，*Kunming University of Science and Technology*，*Kunming* 650500，*China*)

* *Corresponding author*，*E-mail*：*zhangyinhui@kust.edu.cn*

Abstract：Video-Super Resolution (VSR) aims to reconstruct low-resolution video frame sequences into high-resolution video frame sequences. Compared with single image super-resolution，VSR usually relies on the height-dependent information of neighboring frames to reconstruct the current frame because of the added information of temporal dimension. How to align adjacent frames and obtain highly correlated information between frames is the key issue of VSR task. In this paper，the VSR task is divided into three stages：deblurring，alignment，and reconstruction. In the deblurring stage，the current frame is pre-aligned with adjacent frames to obtain feature information highly related to the current frame，and the details of the current frame are enhanced to achieve more feature information extraction in the initial stage. In the alignment stage，the highly correlated information in adjacent frames is used to further strengthen the feature information in the current frame by performing a secondary alignment operation on the input features. In the reconstruction stage，raw low-resolution frames are aggregated to provide more feature information at the end of the network. In this paper，we use Multi-Layer Perceptron (MLP) instead of the traditional convo-

收稿日期：2022-12-14；**修订日期**：2023-01-13.
基金项目：国家自然科学基金资助项目(No. 52065035，No. 62061022，No. 62171206)

lution operation to construct a feature extraction module, and also perform a secondary alignment of the generated feature information to refine the image features to obtain better video frame reconstruction results. The experimental results show that the proposed algorithm achieves a higher accuracy of video frame sequence reconstruction on a variety of publicly available datasets while achieving a lower number of network parameters and a more coherent video sequence reconstruction performance.

Key words: computer vision; video super-resolution; multi-layer perceptron; attention mechanism; optical flow; frame alignment

1　引　言

视频是传递信息的重要媒介之一，对低分辨率（Low Resolution，LR）视频进行超分辨率重建可以有效提高图像和视频的清晰度。自从Dong等人[1]首次将卷积神经网络（Convolutional Neural Network，CNN）引入图像超分（Single Image Super-Resolution，SISR）领域后，大量基于CNN架构的优秀SISR网络[2-7]便得到不断衍生并取得了优异的成果。但大部分SISR网络仅对视频帧序列进行逐帧重建，可能会导致输出结果产生伪影或干扰，无法保证重建视频序列的连续性[8]。

Kappeler等人[9]在SRCNN的基础上首次提出了一种基于CNN的视频超分辨率网络，此后开始涌现出大量基于CNN架构的VSR网络[10-12]。对于一段视频序列而言，VSR将SISR一次处理单帧图像扩展到一次性处理多个连续帧。由于相邻帧中可能包含恢复当前帧的高度相关特征信息，合理利用这些高度相关信息可更好的对当前帧进行重建。但在一个视频帧序列中，同一特征在前后帧中的位置可能不同，如何准确的对齐相邻帧与当前帧之间的高度相关特征便成为了VSR任务的核心问题。

有学者采用运动估计和运动补偿[10,13-15]的方法提取帧与帧间的运动信息，并根据帧间运动信息进行帧与帧之间的图像变换操作，使相邻帧对齐[8]。这类方法多以光流法进行操作，但仅依靠光流法对相邻帧图像进行对齐会因估计误差导致对齐后的帧存在伪影缺陷[16]。有学者利用可变卷积[12,17-18]计算两帧之间的偏移量，以实现帧间相关信息提取。或者凭借3D卷积[19-21]对输入帧序列在时-空域（Spatio-temporal domain）中进行处理，通过提取时间信息处理帧间相关性。但相比于二维卷积而言，可变卷积和3D卷积计算复杂度相对较高，限制了它们在实时视频超分辨率任务中的应用[8]。同时也有利用循环卷积神经网络[22-24]对视频中包含的时空信息进行建模，以实现相邻帧中相似特征提取的操作。但传统基于循环卷积神经网络的方法难以训练，甚至出现梯度消失的问题，尤其是当输入序列的长度太大时，这类方法可能无法获得很好的性能[8]。VSR任务潜在的复杂性和网络框架不同的设计方法在体现各自优势的前提下，也为具体实施和扩展现有的方法带来了困难，阻碍了可重复性和公平的[10]。

本文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。通过对输入视频帧序列进行预对齐，可取得更优异的帧序列重建表现。在后续与其余VSR网络的对比中，MSVSR同样获取了最佳的重建效果。在下文的消融实验部分中，也证明了对视频帧序列进行预对齐操作，可在视频帧序列重建时提升0.01 dB的精度。鉴于同一帧图像在不同尺度下显示的特征信息不同，本文在去模糊阶段提出了一种编码-解码（Encoder-Decoder）架构的多尺度去模糊模块，更加精细化的提取不同尺度的细节特征和全局特征。引入Shift-MLP[25]代替传统卷积的操作方式可以对同一帧图像的不同区域的特征信息进行交互，而本文设计的特征融合块可以对原始输入图像及深层特征进行有效融合，以便进一步获得图像更为详细的特征细节。

2　本文算法

本文提出的MSVSR架构如图1所示。在去

模糊阶段对输入视频帧序列进行预对齐后，可以通过提取相邻帧与当前帧之间的高度相关信息实现图像特征信息增强；将 Shift-MLP 作为后续特征提取的主要工具，可以使 MSVSR 模型在网络参数量更少的情况下获得更优的视频帧序列重建精度。对提取后的特征信息进行二次对齐操作后，本文对重建阶段进行了轻量化设计，将深层特征及原始输入图像进行特征聚合，通过弥补超分辨率流程中损失的特征信息以达到更优的视频帧序列重建效果。

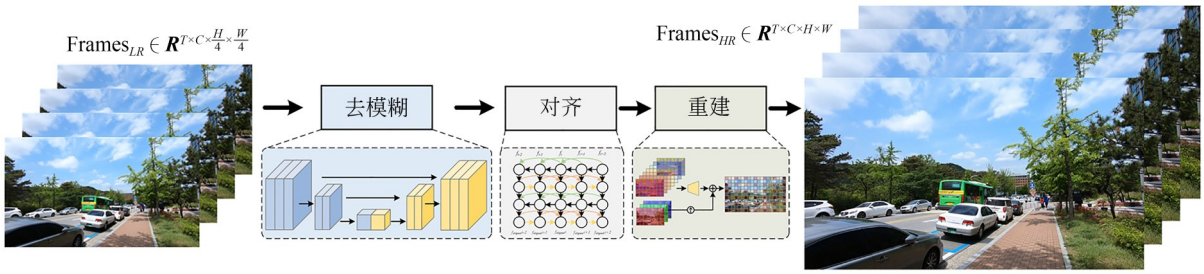


图 1　MSVSR 网络架构

Fig. 1　Architecture of MSVSR

2.1　视频帧去模糊

同一图像在不同尺度下体现出的特征信息不同。因此，本文在去模糊阶段设计了一种编码-解码架构的多尺度去模糊模块。通过在模块中结合预对齐模块及特征融合块，以达到更好的图像特征提取效果。去模糊阶段架构如图 2 所示。

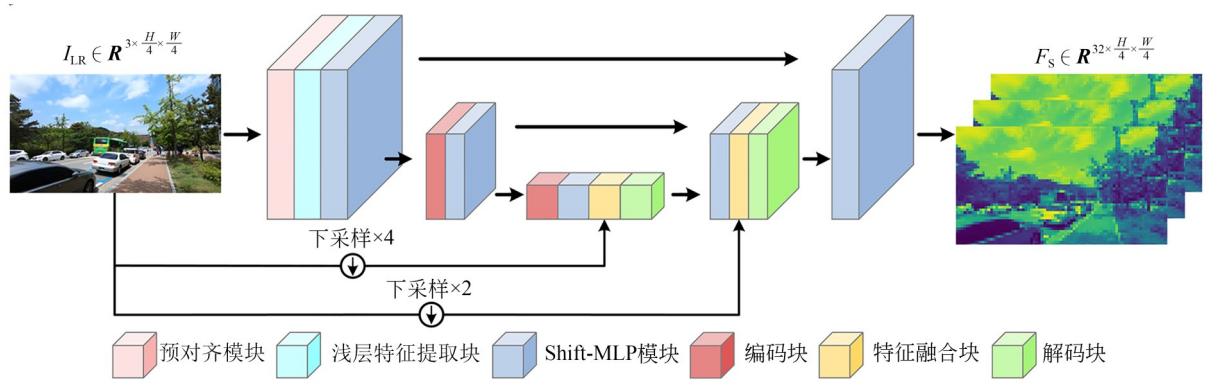


图 2　去模糊阶段

Fig. 2　Deblurring stage

2.1.1　特征预对齐

相较于 SISR 仅从单张图像中进行特征提取，结合邻近帧进行特征提取的 VSR 可更有效的提取更多细节信息。本文在 VSR 任务初始阶段对当前帧与相邻帧进行对齐操作，以结合相邻帧中的高度相关信息从而对当前帧中特征进行增强，更丰富的特征细节可获取更佳的视频帧重建表现。

鉴于相邻帧对齐操作的优势，本文采用对相邻帧进行运动估计和运动补偿的方法实现多尺度去模糊模块中的预对齐操作，凭借 SpyNet[26] 光流法实现相邻帧的特征对齐。光流法的最大优势在于光流法将视频序列中目标的位移映射到一组特征图上，再将经光流估计的特征图输入到后续网络中进行运算。相较于其余对齐方法，采用光流法作为相邻帧对齐操作可降低部分计算开销。

如图 3 所示，本文在多尺度去模糊模块中利用光流法 SpyNet 对相邻帧进行运动估计，获取当前帧的前向传播光流 $flow_i^{\text{forward}}$ 以及反向传播光流 $flow_i^{\text{backward}}$。将获取的 $flow_i^{\text{forward}}$ 和 $flow_i^{\text{backward}}$ 与原始帧 $frame_i$ 输入至光流对齐块进行对齐操

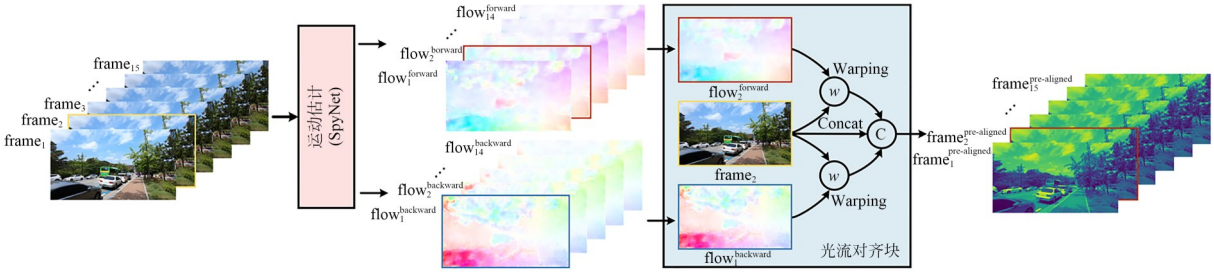


图 3　预对齐模块

Fig. 3　Pre-aligned module

作，可以得到以下的预对齐帧 $flow_i^{\text{pre-aligned}}$：

$$flow_i^{\{\text{backward, forward}\}} = SpyNet(I_i, I_{i\pm1}), \quad (1)$$

其中：I_i, I_{i+1} 为输入的相邻两帧图像，$SpyNet(\cdot)$ 为利用 SpyNet 对相邻帧进行光流运算。预对齐图像 $I_{\text{pre-aligned}}$ 和光流对齐块 $FA(\cdot)$ 可分别表示为：

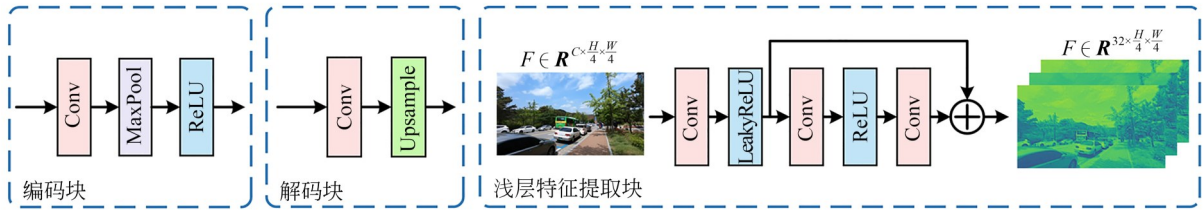
$$I_{\text{pre-aligned}} = FA\left(I_i, flow_i^{\text{forward}}, flow_{i-1}^{\text{backward}}\right), \quad (2)$$

$$FA = C\left(I_i, w\left(flow_i^{\text{forward}}, I_i\right), w\left(flow_{i-1}^{\text{backward}}, I_i\right)\right), \quad (3)$$

其中：$w(\cdot)$ 为 Wraping 操作，$C(\cdot)$ 为 Concat 操作。

2.1.2　特征提取

经过对目标帧进行预对齐操作后，本文将对齐后的帧进行进一步的特征提取操作。本文利用浅层特征提取块（Shallow Feature Extraction Block，SFEB）对目标帧进行浅层特征提取，SFEB 架构细节如图 4(c)所示。由于每帧图像在不同尺度下所蕴含的特征信息不同，因此在去模糊阶段中，本文利用编码块和解码块对图像进行尺度变化，以便提取图像在不同尺度下的特征信息，编码块和解码块的架构细节如图 4(a)和图 4(b)所示。



(a) 编码块　(b) 解码块　(c) 浅层特征提取块

(a) Encoder block　(b) Decoder block　(c) Shallow feature extraction block

图 4　去模糊阶段中的子模块

Fig. 4　Sub-module in deblurring stage

图像中未被遮挡特征可能包含恢复遮挡处的重要特征。为解决视频帧序列中出现目标局部被遮挡而无法精细重建特征的情况，本文除在时域内进行相邻帧对齐外，同时对单帧图像进行不同区域的特征交互处理。CNN 需要依赖卷积核的移动来捕获图像中的目标特征，尤其是图像中的两个特征距离较远时，CNN 受限于其有限的感受野可能会导致网络建模困难。利用多层感知机（Multi-Layer Perceptron，MLP）捕捉两目标物体时，计算量不会随着距离的增加而增大，这样可以很好地解决长距离依赖问题。通过对图像进行多方向上的滑动操作，可将图像中不同区域的信息进行交互，对交互后的图像数据进行计算后可实现跨区域间像素的信息交流，从而通过间接扩大了模型的感受野的方式提高信息的有效使用性。因此，本文引入文献[25]中的 Shift-MLP 操作，并融合卷积及 LayerNorm 等操作来构建如图 5 所示的深层特征提取核心模块 Shift-MLP。其具体流程可表示为：

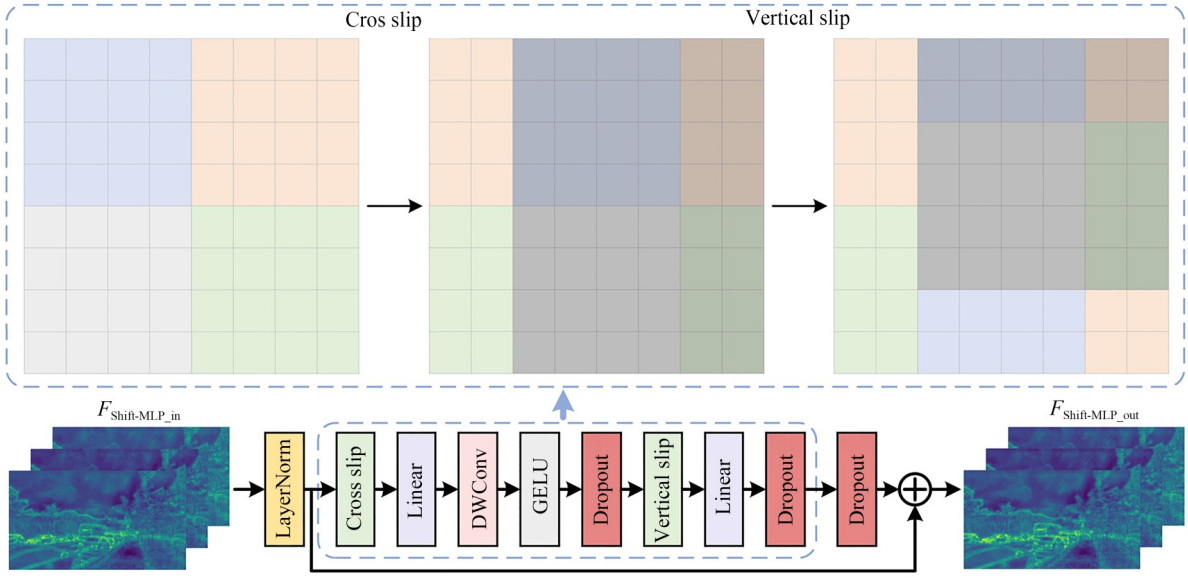


图 5　Shift-MLP 模块

Fig. 5　Shift-MLP module

$$F_{DF}=Dropout\left(ShiftMLP\left(LN\left(F_{\text{pre-aligned}}\right)\right)\right)+LN\left(F_{\text{pre-aligned}}\right),\tag{4}$$

其中：$LN(\cdot)$为 LayerNorm 操作，$F_{\text{pre-aligned}}$为经过前文预对齐后的帧。而 Shift-MLP 可表示为：

$$X_{\text{slip\_}w}=Slip_w(X);\ T_w=Tokenize\left(X_{\text{slip\_}w}\right),\tag{5}$$

$$Y=Dropout\left(GELU\left(DWConv\left(MLP(T_w)\right)\right)\right),\tag{6}$$

$$Y_{\text{slip\_}h}=Slip_h(Y);\ T_h=Tokenize\left(Y_{\text{slip\_}h}\right),\tag{7}$$

$$Y=Dropout\left(MLP(T_h)\right),\tag{8}$$

其中：$Slip_w(\cdot)$为水平方向滑动（Cross Slip），$Slip_h(\cdot)$为垂直方向滑动（Vertical Slip），$DWConv(\cdot)$为深度卷积（Depth-Wise Convolution，DWConv），$Tokenize(\cdot)$为对滑移后的图像进行编码操作以便后续计算。在 Shift-MLP 操作中，本文采用深度卷积代替传统卷积，以便在视频帧序列重建精度类似的情况下取得更低的模型参数量表现。

为有效融合不同尺度下图像特征信息，本文引入文献[27]中的自监督模块，对其进行优化后作为特征融合模块（Feature Fusion Block，FFB）。FFB 的架构细节如图 6 所示。本文将 Shift-MLP 模块输出的特征图$F_{\text{Shift-MLP_out}}$与原始低分辨率视频帧I_{LR}进行聚合，利用I_{LR}中丰富的特征细节信息弥补在特征提取过程中损失的特征细节。由图 6 可以看出，相较于原始特征图$F_{\text{Shift-MLP_out}}$，融合后I_{LR}后的的特征F_{fusion}在边缘细节上已变得更加清晰。

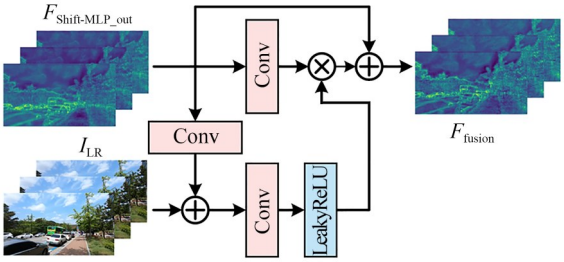


图 6　特征融合块

Fig. 6　Feature fusion block

2.2　视频帧对齐

相较于 SISR 任务，VSR 任务难点便在于帧对齐操作。通过对当前帧与相邻帧进行对齐操作，交互并提取当前帧与相邻帧之间的高度相关信息，强化当前帧缺失的特征细节。可提升视频帧序列重建的精度及连贯性。现有 VSR 网络中的采用的对齐模块结构，大致可分为如图 7 所示的 4 种[28]。

（1）多帧融合[15,29-30]：输入一个视频序列，将整个视频序列视为多个独立的子过程。这些子过程在时间上不相关，可独立处理。因此，该种

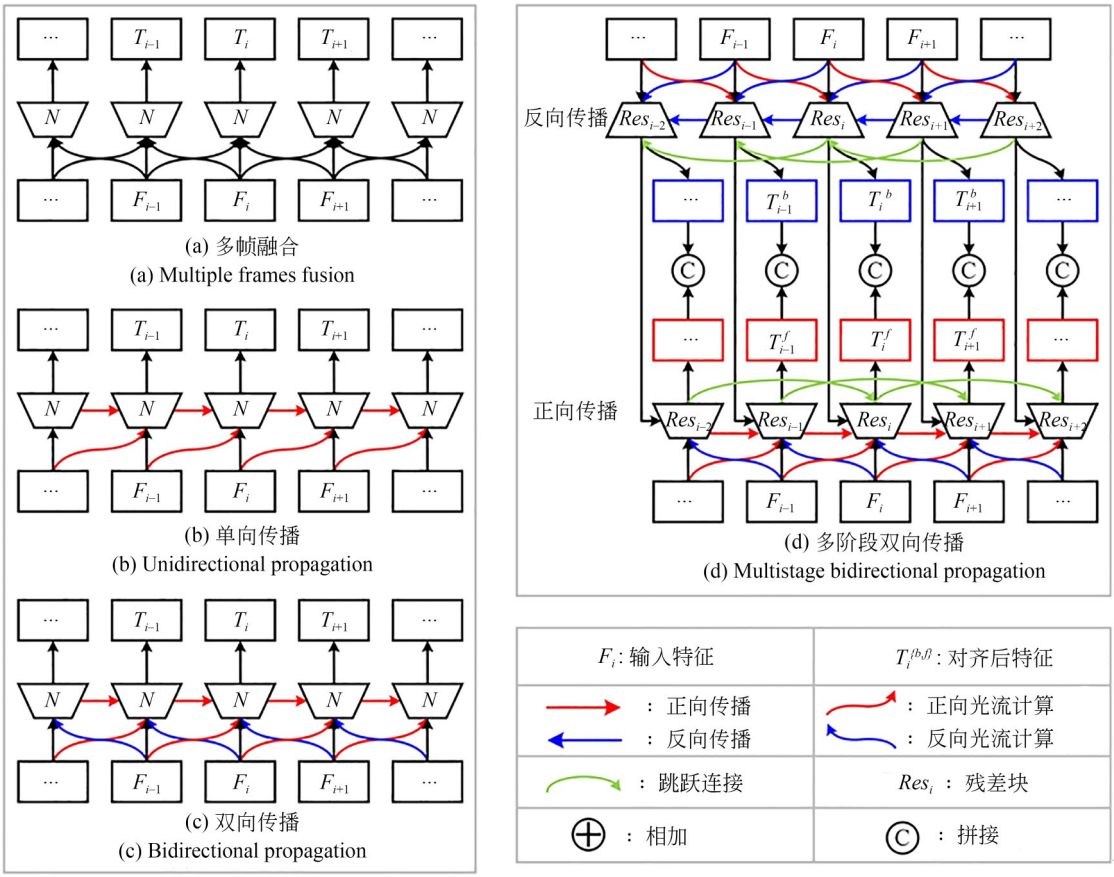


图 7　对齐阶段[28]

Fig. 7　Alignment stage[28]

方法享有并行计算的优势[31]。然而这种迭代方法忽略了先前估计的 SR 输出。而多帧融合(Multiple frames fusion)直接将未经对齐的特征聚合后还原，传播特征在空间上和输入图像不一致导致方法难以取得优异的性能[28]。

(2)单向传播[32-33]：该类方法利用相邻帧进行正向光流计算，利用计算所得的光流与当前帧进行进一步处理，处理后的结果沿正向传播逐步传递。在单向传播(Unidirectional propagation)的情况下，视频序列中不同帧所获取的信息是不平衡的。具体来说，第一帧除自身信息外无法获取来自视频序列的信息，仅有最后一帧能够获取来自整个序列的信息。因此，在视频序列中较早的帧会出现次优结果[10]。

(3)双向传播[24,34]：与单向传播基本类似，双向传播(Bidirectional propagation)利用相邻帧分别计算正向光流和反向光流。聚合正向光流、反向光流和当前帧，作为下一个模块的输入，在将计算所得沿正向传播逐步传递。通过该方法可使视频帧序列中每一帧均可获得相邻帧之间的特征信息。由于该种方法仅计算相邻两帧之间的光流信息，因此该种方法在重建当前帧时，获取到的相邻帧相关特征是有限的。

(4)多阶段双向传播[11,28]：多阶段双向传播(Multistage bidirectional propagation)分别对输入的正向光流、反向光流和当前帧进行正向传播和反向传播，以期获取远距离帧的高度相关特征。

为对齐经过去模糊后的输入特征信息，本文采用与文献[11]一致的多阶段双向传播对齐方法，利用前文预对齐阶段时计算的视频帧序列正向光流 $flow_i^{\text{forward}}$，反向光流 $flow_i^{\text{backward}}$，对去模糊阶段输出的特征图序列进行进一步对齐。$flow_i^{\{\text{backward, forward}\}}$ 计算过程公式(9)所示。本文将计算所得的光流序列与特征序列进行聚合，在将聚合后的特征序列输入至残差块中进行下一步的计算。

$$flow_i^{\{\text{backward, forward}\}} = SpyNet(I_i, I_{i\pm1}), \quad (9)$$

$$F_{\text{hybrid}} = F_i + flow_i^{\text{forward}} + flow_i^{\text{backward}}. \quad (10)$$

由于视频序列中前后帧均有与当前帧高度相关的特征信息，因此本文对残差块的输出分别进行正向传播和反向传播，同时加入跳跃连接以更好的捕获更远距离帧中的特征信息。在正向传播和反向传播的过程中，依次交互两者数据可以更好的获取视频帧序列中与当前帧相似的特征信息。

$$T_i^{\{b,f\}} = \boldsymbol{R}_i^{\{b,f\}}\left(F_{\text{hybrid}}\right), \quad (11)$$

其中，残差块操作$\boldsymbol{R}(\cdot)$可表示为：

$$\boldsymbol{R}_i^{\{b,f\}} = \boldsymbol{C}\left(\boldsymbol{D}\left(T_{i-1}^{\{b,f\}}, T_{i-2}^{\{b,f\}}\right), F_{\text{hybrid}}\right). \quad (12)$$

$T_{i-1}^{\{b,f\}}, T_{i-2}^{\{b,f\}}$分别为反向或正向传播过程中前序残差块输出的结果，$\boldsymbol{D}(\cdot)$为可变卷积操作。正向传播和反向传播依次交替数据可表示为：

$$\boldsymbol{R}_i^{f_j} = \boldsymbol{R}_i^{f_{j-1}} + \boldsymbol{R}_i^{b_{j-2}}, \quad (13)$$

$$\boldsymbol{R}_i^{b_j} = \boldsymbol{R}_i^{b_j} + \boldsymbol{R}_i^{f_{j-1}}, \quad (14)$$

$$\boldsymbol{R}_i^{f_{j+1}} = \boldsymbol{R}_i^{f_{j+1}} + \boldsymbol{R}_i^{b_j}, \quad (15)$$

其中：$\boldsymbol{R}_i^{f_{j-1}}$中的$f_{i-1}$表示第$j-1$次传播为正向传播，$b_j$表示第$j$次传播为反向传播。

将反向传播和正向传播输出后的特征图进行拼接后即可得到对齐帧F_i^{align}，其公式表示为：

$$F_i^{\text{align}} = \boldsymbol{C}\left(T_i^b, T_i^f\right). \quad (16)$$

通过对输入的特征图进行二次对齐操作，可利用相邻帧中的高度相关信息对当前帧进行特征增强，从而弥补在多次特征提取中损失的特征细节。

2.3 视频帧重建

在图 8 的图像重建模块中，本文聚合了原始低分辨率帧以及经过对齐模块后的对齐帧F_{aligned}后，将其输入图 4 中的 SFEB 进行进一步的特征提取和恢复操作。对二次对齐特征F_{aligned}进行进一步重建后，本文将低分辨率原始帧I_{LR}进行上采样操作并与重建后的特征帧进行残差连接，以在网络的末端赋予图像更多的特征细节。

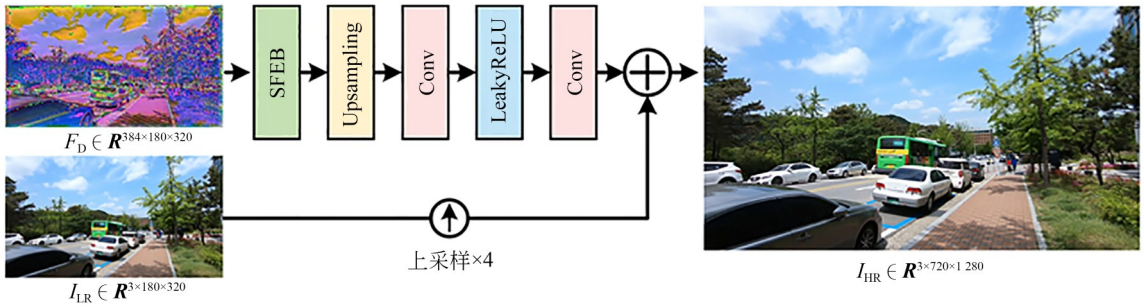


图 8　重建阶段

Fig. 8　Reconstruction stage

2.4 训练细节

本文采用与 BasicVSR[10]一致的预训练参数设置，将预训练的 SpyNet[26]作为运动估计方法，初始学习率设置为2×10^{-4}，采取余弦退火(Cosine Annealing scheme)[35]学习率衰减策略，以及 Adam[36]优化器进行网络模型的参数优化，利用 NVIDIA GeForce RTX 3090Ti 对 MSVSR 进行 60 万次的迭代训练。本文采用 Charbonnier Loss 作为训练时的网络损失函数，Charbonnier 损失函数定义如公式(17)所示。

$$L_{\text{Charbonnier}} = \sqrt{\| I - I_{GT} \|^2 + \varepsilon^2}, \quad (17)$$

其中：I为重建后的视频帧，I_{GT}为对应的原始视频帧(Ground-Truth，GT)，ε为取值等于1×10^{-3}的常量。与 L1 损失相比，由于增加了一个正则项ε，因此接近零点的值的梯度由于ε的存在，梯度不会太小，可避免在训练过程中产生梯度消失的情况。

3 实验结果与分析

3.1 实验结果

本文在公开数据集 REDS[37]上进行 BI(Bicubic)任务的训练，并将训练所得的权重在 Vid4[38]的 BI 数据集和 REDS4(REDS 训练集中的 000，011，015，020 文件)上进行测试。同时本文在

Vimeo90K[14]数据集上进行BD(Blur and Downsampling)任务的训练，并将测试结果在Vimeo90K-T[14](Vimeo-90K的测试集)及Vid4的BD数据集上进行测试。将本文算法与选取的VESPCN[13]，TOF[14]，DUF[19]，FRVSR[33]，SPMC[15]，EDVR[12]，EDVR-M[12]，MuCAN[39]，TGA[40]，RLSP[24]，RSDN[32]，RRN[41]，RBPN[42]，PFNL[43]，BasicVSR[10]进行VSR任务的定量对比结果如表1所示(除REDS4在RGB通道上计算外，所有结果均在Y通道上计算，表中的加粗部分表示最佳性能，运行时间按低分辨率图像尺寸为180×320计算，空白条目对应于以往工作中未报道的结果)。

网络参数量和图像重建精度对比参见图9所示。对比网络在公开数据集上的视频帧重建结果来源于BaiscVSR[10]论文中提供的数据。

从表1中可以得出，MSVSR在BI与BD任务中都取得了比BasicVSR更优的图像重建效果，而MSVSR的参数量相比BasicVSR却少了近一半。而相比EDVR-M(EDVR轻量化版本)，MSVSR在参数量和EDVR-M相近的情况下，在REDS4数据集上的视频帧序列重建精度比EDVR-M高了1.11 dB。

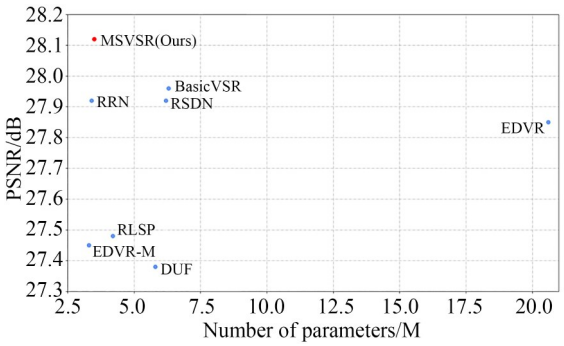


图9　在Vid4数据集上进行BD任务的性能和模型复杂度对比

Fig. 9　Performance and model complexity comparison on Vid4 dataset for BD task

表1　定量比较[10]

Tab. 1　Quantitative comparison[10]

对比网络	参数量/M	运行时间/ms	BI(PSNR/SSIM)		BD(PSNR/SSIM)	
			REDS4	Vid4	Vid4	Vimeo-90K-T
Bicubic	—	—	26.14/0.729 2	23.78/0.634 7	21.80/0.524 6	31.30/0.868 7
VESPCN[13]	—	—	—	25.35/0.755 7	—	—
TOF[14]	—	—	27.98/0.799 0	25.89/0.765 1	—	34.62/0.921 2
DUF[19]	5.8	974	28.63/0.825 1	—	27.38/0.832 9	36.87/0.944 7
FRVSR[33]	5.1	137	—	—	26.69/0.810 3	35.64/0.931 9
SPMC[15]	—	—	—	25.88/0.775 2	—	—
EDVR[12]	20.6	378	31.09/0.880 0	27.35/0.826 4	27.85/0.850 3	37.81/0.952 3
EDVR-M[12]	3.3	118	30.53/0.869 9	27.10/0.818 6	27.45/0.840 6	37.33/0.948 4
MuCAN[39]	—	—	30.88/0.875 0	—	—	—
TGA[40]	5.8	—	30.88/0.875 0	—	—	—
RLSP[24]	4.2	49	—	—	27.48/0.838 8	36.49/0.940 3
RSDN[32]	6.2	94	—	—	27.92/0.850 5	37.23/0.947 1
RRN[41]	3.4	**45**	—	—	27.92/0.850 5	37.23/0.947 1
PFNL[43]	**3.0**	295	29.63/0.850 2	26.73/0.802 9	27.16/0.833 5	—
RBPN[42]	12.2	1 507	30.09/0.859 0	27.12/0.818 0	—	37.20/0.945 8
BaiscVSR[10]	6.3	63	31.42/0.890 9	27.24/0.825 1	27.96/0.855 3	37.53/0.949 8
MSVSR	3.5	245	**31.64/0.895 8**	**27.41/0.833 4**	**28.12/0.858 2**	**37.70/0.951 4**

由图10可得，MSVSR在REDS4数据集上取得了更好的图像重建效果。尤其是在车牌重建细节上，MSVSR展现出了更锐利数字边缘。同时如图11所示，MSVSR在Vid4数据集上重建出的视频帧序列在观感上也优于其余对比网络。例如，在Calendar文件夹下，MSVSR重建

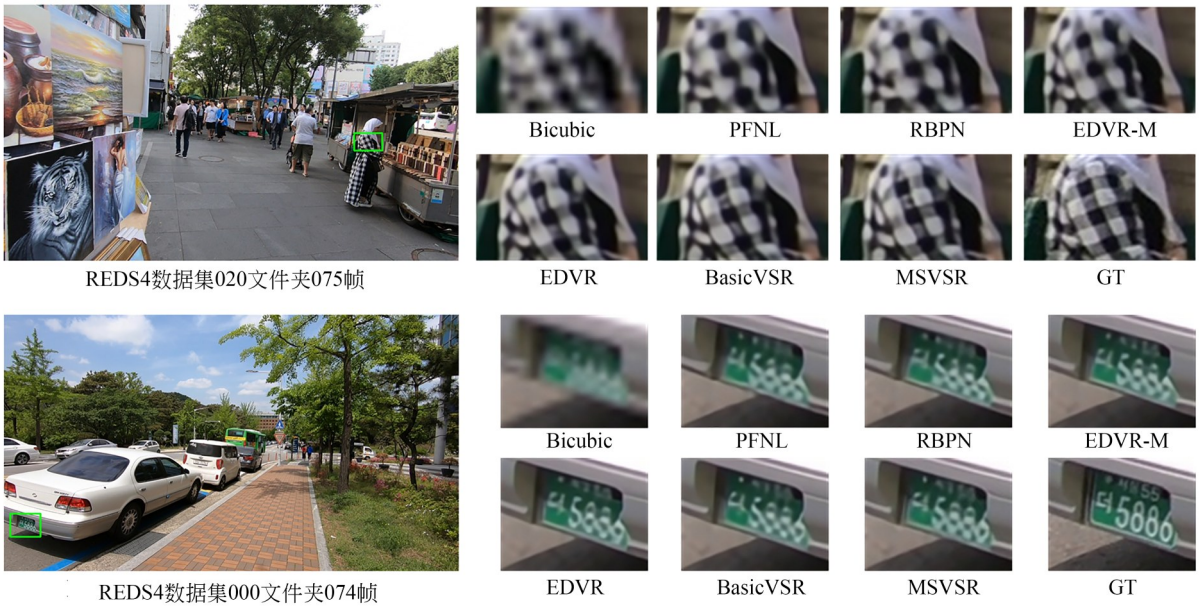


图 10　REDS4 数据集上的定量对比[10]

Fig. 10　Qualitative comparison on REDS4[10]

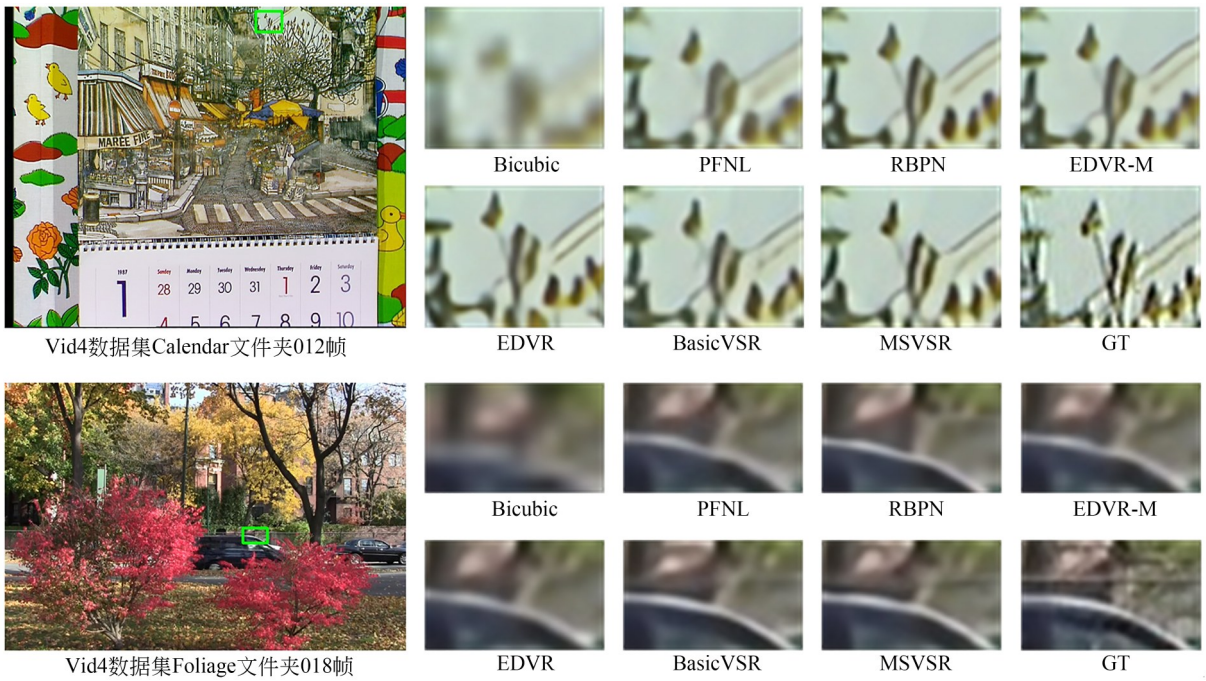


图 11　Vid4 数据集上的定量对比[10]

Fig. 11　Qualitative comparison on Vid4[10]

后的竹枝边缘取得了更锐利的观感与 GT 数据最为接近，而其余对比网络重建出的竹枝边缘已模糊不清。并且在 Foliage 文件夹下，MSVSR 重建出的左上角树木边缘也更加锐利，且细节特征也更加丰富。如图 12 所示，在 Vimeo-K-T 数据集的 00010/943/010 文件下的图像数据重建结果中，仅有 MSVSR 还原出了字母 O 左上角的一处缺口。并且其余对比网络均未重建出 OIL 三字母边缘黑边细节，仅有 MSVSR 重建出了与 GT 图像最为接近的效果。图 10～图 12 中对比网络重建的结果来源于论文 BasicVSR[10]。

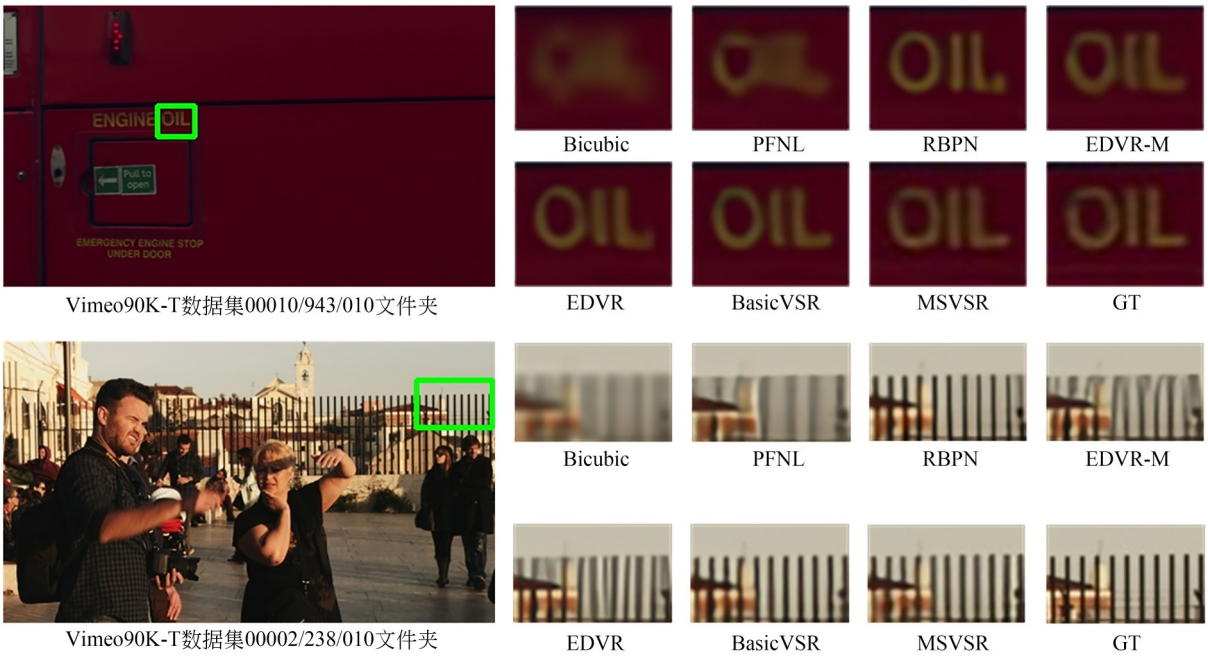


图 12　Vimeo-90K-T 数据集上的定量对比[10]

Fig. 12　Qualitative comparison on Vimeo-90K-T[10]

3.2　消融试验

为验证特征融合块(Feature Fusion Block, FFB)和预对齐模块(Pre-aligned module)的有效性,本文对这两个模块进行了独立的消融实验,并将结果列举于表 2 中。从表中可以得出,利用特征融合块融合深层特征和原始图像细节特征能够明显提升图像的重建质量。由于增加了预对齐模块,输入帧序列在初期便可获得相邻帧的高度相关信息,更多的特征信息可在后续特征处理阶段取得更好的特征重建结构

表 2　在 Vid4 数据集中 BD 任务上的消融实验

Tab. 2　Ablation tests on Vid4 dataset for BD task

FFB	Pre-aligned	PSNR	SSIM
√	√	28.12	0.858 2
√	×	28.11	0.858 8
×	√	28.10	0.858 6
×	×	28.08	0.856 9

为验证前文所述的“相比于传统卷积,DWConv 可以在视频帧序列重建精度类似的情况下取得更低的模型参数量表现”,本文对去模糊阶段中的 Shift-MLP 模块进行定量的客观性比较,在此处仅将 Shift-MLP 模块中的传统卷积换为 DWConv。表 3 的实验结果与前文表述的一致性表明,传统卷积替换为 DWConv 后,模型参数量下降了 0.38M(9.7%)而 PSNR 结果损失不足 1%。

表 3　在 Vid4 数据集上 BD 任务中通过改变不同卷积的定量对比

Tab. 3　Quantitative comparison of different convolution decomposition approaches on Vid4 dataset for BD task

Method	Params	PSNR	SSIM
Shift-MLP(DWConv)	3.53 M	28.12	0.858 2
Shift-MLP(Conv)	3.91 M	28.14	0.859 9

3.2　连贯性对比

为证明两次对齐方法可取得更好的视频帧序列重建连贯性,本文在 BI 任务下特意选取 GT 视频帧数据,以及经 MSVSR,BasicVSR,EDVR 重建后的 Vid4 数据集 Calenda 文件夹下的视频帧进行可视化比较。本文选取一列红色虚线对视频帧序列进行剖切,观察同一区域随时间的变化。重建视频帧序列在时间剖面上的连贯性比较如图 12 所示。通过主观性比较可知,EDVR 重建出的视频数据在连贯性较差,帧与帧之间出现抖动现象,而 MSVSR 取得了与 GT 最为相近视

频帧序列重建的连贯性表现。

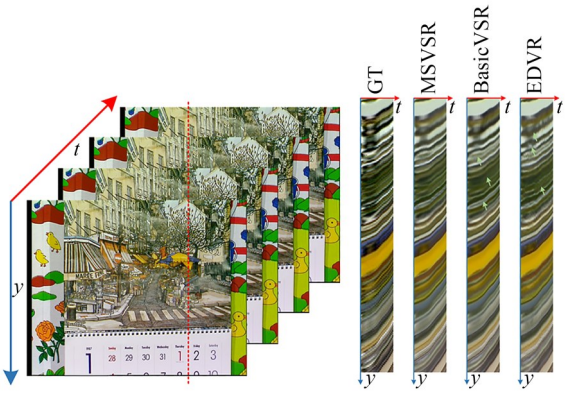


图 13　时间剖面的比较

Fig. 13　Comparison of temporal profile

4　结　论

本文提出了一种多阶段帧对齐的视频超分辨率网络，通过将视频超分任务分为视频去模糊，视频帧对齐和重建三个阶段实现清晰的视频重建。通过在去模糊阶段对输入视频帧序列进行预对齐获得相邻帧序列高度相关特征信息。利用特征融合块将深度特征与原始图像的特征细节进行融合，获取了更为优异的特征重建质量。本文利用Shift-MLP结合深度卷积作为特征提取模块，相较于对比的VSR网络，MSVSR在网络参数量更低的情况下，重建得到的视频帧序列在PSNR/SSIM值的评估比较中均取得了最好的结果。

更清晰的视频帧序列重建效果，更连贯的视频帧重建表现，更小的模型量参数，这使得MSVSR可以更快的落实到实际工业应用中。以结构体视觉振动位移测量[44-45]为例，通过对输入视频帧序列进行更精细地重建，可回归出更为贴近真实场景的结构体位移信息。相较于对比网络，MSVSR在视频帧序列重建精度更高的情况下网络参数量仅为对比网络的一半。在后期的研究中，作者团队将对MSVSR进行进一步优化，争取在参数量大幅降低的情况下，视频帧序列重建精度取得更优秀的表现。

参考文献：

[1] DONG C, LOY C C, HE K M, *et al*. *Learning a Deep Convolutional Network for Image Super-Resolution*[M]. Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 184-199.

[2] LIM B, SON S, KIM H, *et al*. Enhanced deep residual networks for single image super-resolution[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*). 21-26, 2017, *Honolulu*, *HI*, *USA*. IEEE, 2017: 1132-1140.

[3] ZHANG Y L, LI K P, LI K, *et al*. *Image Super-Resolution Using Very Deep Residual Channel Attention Networks*[M]. Computer Vision - ECCV 2018. Cham: Springer International Publishing, 2018: 294-310.

[4] SHI W Z, CABALLERO J, HUSZÁR F, *et al*. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). 27-30, 2016, *Las Vegas*, *NV*, *USA*. IEEE, 2016: 1874-1883.

[5] 蔡体健，彭潇雨，石亚鹏，等. 通道注意力与残差级联的图像超分辨率重建[J]. 光学 精密工程，2021，29(1)：142-151.
CAI T J, PENG X Y, SHI Y P, *et al*. Channel attention and residual concatenation network for image super-resolution[J]. *Opt. Precision Eng.*, 2021, 29(1): 142-151. (in Chinese)

[6] 程德强，赵佳敏，寇旗旗，等. 多尺度密集特征融合的图像超分辨率重建[J]. 光学 精密工程，2022，30(20)：2489-2500.
CHENG D Q, ZHAO J M, KOU Q Q, *et al*. Multi-scale dense feature fusion network for image super-resolution[J]. *Opt. Precision Eng.*, 2022, 30(20): 2489-2500. (in Chinese)

[7] 耿铭昆，吴凡路，王栋. 轻量化火星遥感影像超分辨率重建网络[J]. 光学 精密工程，2022，30(12)：1487-1498.
GENG M K, WU F L, WANG D. Lightweight Mars remote sensing image super-resolution reconstruction network[J]. *Opt. Precision Eng.*, 2022, 30(12): 1487-1498. (in Chinese)

[8] LIU H Y, RUAN Z B, ZHAO P, *et al*. Video super-resolution based on deep learning: a comprehensive survey [J]. *Artificial Intelligence Review*, 2022, 55(8): 5981-6035.

[9] KAPPELER A, YOO S, DAI Q Q, *et al*. Video super-resolution with convolutional neural networks [J]. *IEEE Transactions on Computational Imaging*, 2016, 2(2): 109-122.

[10] CHAN K C K, WANG X T, YU K, *et al*. BasicVSR: the search for essential components in video super-resolution and beyond [C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 20-25, 2021, *Nashville*, *TN*, *USA*. IEEE, 2021: 4945-4954.

[11] 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*). 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 5962-5971.

[12] 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*). 16-17, 2019, *Long Beach*, *CA*, *USA*. IEEE, 2020: 1954-1963.

[13] CABALLERO J, LEDIG C, AITKEN A, *et al*. Real-time video super-resolution with spatio-temporal networks and motion compensation [C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). 21-26, 2017, *Honolulu*, *HI*, *USA*. IEEE, 2017: 2848-2857.

[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] 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*). 22-29, 2017, *Venice*, *Italy*. IEEE, 2017: 4482-4490.

[16] YAN Q S, GONG D, SHI Q F, *et al*. Attention-guided network for ghost-free high dynamic range imaging [C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 15-20, 2019, *Long Beach*, *CA*, *USA*. IEEE, 2020: 1751-1760.

[17] KUPYN O, MARTYNIUK T, WU J R, *et al*. DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better [C]. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 27-*November* 2, 2019, *Seoul*, *Korea* (*South*). IEEE, 2020: 8877-8886.

[18] 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*). 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 3357-3366.

[19] 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*. 18-23, 2018, *Salt Lake City*, *UT*, *USA*. IEEE, 2018: 3224-3232.

[20] KIM S Y, LIM J, NA T, *et al*. 3DSRnet: Video Super-Resolution Using 3D Convolutional Neural Networks [EB/OL]. 2018: *arXiv*: 1812.09079. https://arxiv.org/abs/1812.09079

[21] LI S, HE F X, DU B, *et al*. Fast spatio-temporal residual network for video super-resolution [C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 15-20, 2019, *Long Beach*, *CA*, *USA*. IEEE, 2020: 10514-10523.

[22] HUANG Y, WANG W, WANG L. Video super-resolution via bidirectional recurrent convolutional networks [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, 40(4): 1015-1028.

[23] ZHU X B, LI Z Z, ZHANG X Y, *et al*. Residual invertible spatio-temporal network for video super-resolution [J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, 33(1): 5981-5988.

[24] FUOLI D, GU S H, TIMOFTE R. Efficient video super-resolution through recurrent latent space propagation [C]. 2019 *IEEE/CVF International Conference on Computer Vision Workshop* (*ICCVW*). 27-28, 2019, *Seoul*, *Korea* (*South*).

IEEE, 2020: 3476-3485.

[25] VALANARASU J M J, PATEL V M. *UNeXt: MLP-Based Rapid Medical Image Segmentation Network*[M]. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 23-33.

[26] RANJAN A, BLACK M J. Optical flow estimation using a spatial pyramid network[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). *July* 21-26, 2017. *Honolulu, HI.* IEEE, 2017: 4161-4170.

[27] ZAMIR S W, ARORA A, KHAN S, *et al*. Multi-stage progressive image restoration[C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 20-25, 2021. *Nashville, TN, USA*. IEEE, 2021: 14821-14831.

[28] YI P, WANG Z Y, JIANG K, *et al*. Omniscient video super-resolution[C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 10-17, 2021. *Montreal, QC, Canada*. IEEE, 2021: 4429-4438.

[29] WANG Z Y, YI P, JIANG K, *et al*. Multi-memory convolutional neural network for video super-resolution[J]. *IEEE Transactions on Image Processing*, 2019, 28(5): 2530-2544.

[30] YI P, WANG Z Y, JIANG K, *et al*. Multi-temporal ultra dense memory network for video super-resolution[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020, 30(8): 2503-2516.

[31] YI P, WANG Z Y, JIANG K, *et al*. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations[C]. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 27 - *November* 2, 2019, *Seoul, Korea* (*South*). IEEE, 2020: 3106-3115.

[32] ISOBE T, JIA X, GU S H, *et al*. *Video Super-Resolution with Recurrent Structure-Detail Network*[M]. Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 645-660.

[33] SAJJADI MS, VEMULAPALLI R, BROWN M. Frame-recurrent video super-resolution[C]. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018: 6626-6634.

[34] YAN B, LIN C, TAN W. Frame and feature-context video super-resolution[C]. *Proceedings of the AAAI conference on artificial intelligence*, 2019, 33(01): 5597-5604.

[35] LOSHCHILOV I, HUTTER F. SGDR: Stochastic Gradient Descent with Warm Restarts[EB/OL]. 2016: *arXiv*: 1608.03983. https://arxiv.org/abs/1608.03983

[36] KINGMA D P, BA J. Adam: a Method for Stochastic Optimization [EB/OL]. 2014: *arXiv*: 1412.6980. https://arxiv.org/abs/1412.6980

[37] NAH S, BAIK S, HONG S, *et al*. NTIRE 2019 challenge on video deblurring and super-resolution: dataset and study[C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*). 16-17, 2019, *Long Beach, CA, USA*. IEEE, 2020: 1996-2005.

[38] 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.

[39] LI W B, TAO X, GUO T A, *et al*. *MuCAN: Multi-Correspondence Aggregation Network For Video Super-Resolution*[M]. Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 335-351.

[40] ISOBE T, LI S J, JIA X, *et al*. Video super-resolution with temporal group attention[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 13-19, 2020, *Seattle, WA, USA*. IEEE, 2020: 8005-8014.

[41] ISOBE T, ZHU F, WANG S. Revisiting Temporal Modeling for Video Super-Resolution [EB/OL]. 2020: *arXiv*: 2008.05765. https://arxiv.org/abs/2008.05765

[42] HARIS M, SHAKHNAROVICH G, UKITA N. Recurrent back-projection network for video super-resolution[C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 15-20, 2019, *Long Beach, CA, USA*. IEEE, 2020: 3892-3901.

[43] YI P, WANG Z, JIANG K, *et al*. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations[C]. *Proceed-*

ings of the IEEE/CVF international conference on computer vision. 2019: 3106-3115.

[44] YANG R, WANG S, WU X, *et al*. Using lightweight convolutional neural network to track vibration displacement in rotating body video[J]. *Mechanical Systems and Signal Processing*, 2022, 177: 109137.

[45] ZHOU J W, LI H G, ZHANG L, *et al*. Vibration measurement with video processing based on alternating optimization of frequency and phase shifts [J]. *IEEE Transactions on Instrumentation and Measurement*, 2021, 70: 1-13.

作者简介：

王　森(1983－)，男，河南信阳人，博士，副教授，硕士生导师，2007年于郑州轻工业大学获得学士学位，2014年于昆明理工大学获得硕士学位，2017年在昆明理工大学获得博士学位，现为昆明理工大学机电工程学院副教授，主要从事机器视觉、视觉智能感知与测量、故障诊断方面的研究。E-mail：wangsen0401@126.com

祝　阳(1998－)，男，江西鹰潭人，硕士研究生，2021年于温州理工学院获得学士学位，现为昆明理工大学机电工程学院硕士研究生，主要从事计算机视觉中图像复原方面的算法研究。E-mail: zhuyang1023@foxmail.com

通讯作者：

张印辉(1977－)，男，河北衡水人，教授，博士生导师，分别于2000年、2005年西安理工大学获得学士、硕士学位，2010年于昆明理工大学获得博士学位，现为昆明理工大学机电工程学院教授，主要从事计算机视觉中图像分割方面的算法研究。E-mail：zhangyinhui@kust.edu.cn