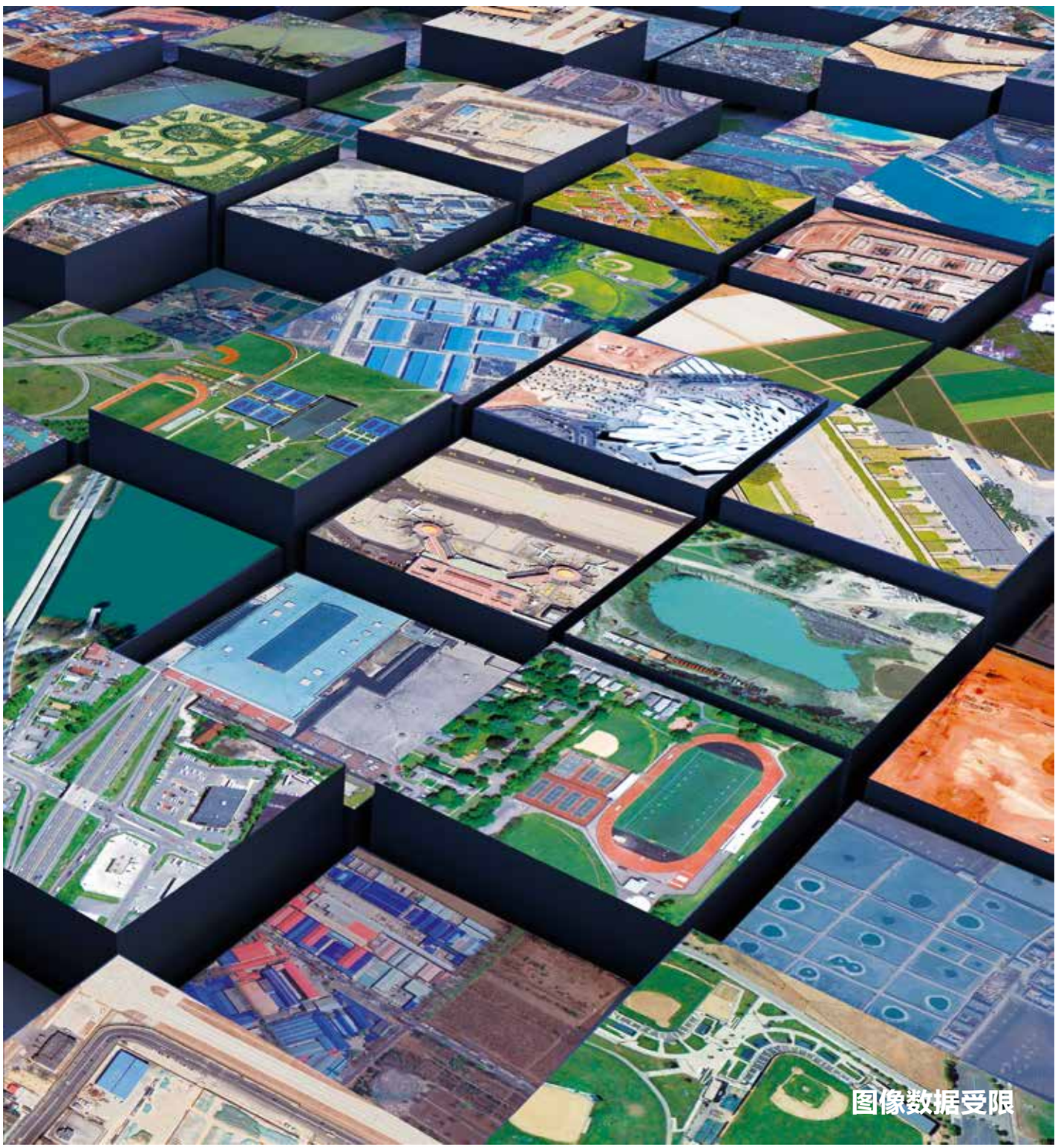
JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2022
10
VOL.27

ISSN1006-8961
CN11-3758/TB



中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第27卷第10期（总第318期）
2022年10月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@aircas.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

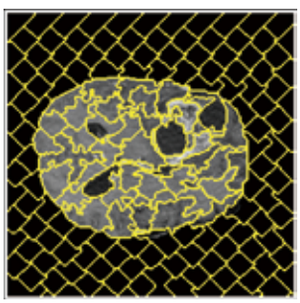
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

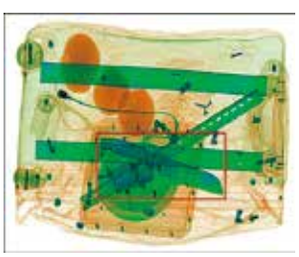
国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2022年10月 第27卷 第10期（总第318期）
中国图象图形学报 Zhongguo Tuxiang Tuxing Xuebao

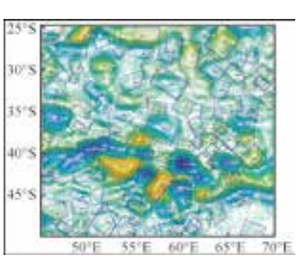
目次



MRI脑肿瘤图像的超像素/体素分割及发展现状(第2897页)



融合多尺度特征与全局上下文信息的X光违禁物品检测(第3043页)

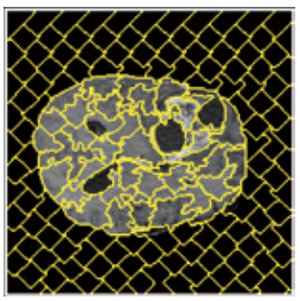


融合多尺度旋转锚机制的海洋中尺度涡自动检测(第3092页)

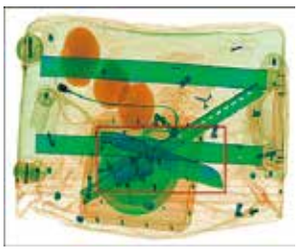
封面素材来自中国科学院空天信息创新研究院二部孙显研究员团队

Volume 27, Number 10
Published October 16, 2022

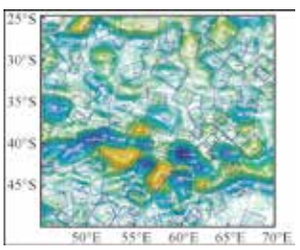
CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



The review of superpixel/voxel segmentation of MRI brain tumor images(P2897)



Integrated multi-scale features and global context in x-ray detection for prohibited items(P3043)



Multi-scale rotating anchor mechanism based automatic detection of ocean mesoscale eddy(P3092)

E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2022)10-2984-10

论文引用格式: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]

轻量级注意力约束对齐网络的视频超分重建

靳雨桐,宋慧慧*,刘青山

南京信息工程大学,江苏省大气环境与装备技术协同创新中心,江苏省大数据分析技术重点实验室,南京　210044

摘　要:**目的**　深度学习在视频超分辨率重建领域表现出优异的性能,本文提出了一种轻量级注意力约束的可变形对齐网络,旨在用一个模型参数少的网络重建出逼真的高分辨率视频帧。**方法**　本文网络由特征提取模块、注意力约束对齐子网络和动态融合分支3部分组成。1)共享权重的特征提取模块在不增加参数量的前提下充分提取输入帧的多尺度语义信息。2)将提取到的特征送入注意力约束对齐子网络中生成具有精准匹配关系的对齐特征。3)将拼接好的对齐特征作为共享条件输入动态融合分支,融合前向神经网络中参考帧的时域对齐特征和原始低分辨率(low-resolution,LR)帧在不同阶段的空间特征。4)通过上采样重建高分辨率(high-resolution,HR)帧。**结果**　实验在两个基准测试数据集(Vid4(Vimeo-90k)和REDS4(realistic and diverse scenes dataset))上进行了定量评估,与较先进的视频超分辨率网络相比,本文方法在图像质量指标峰值信噪比(peak signal to noise ratio,PSNR)和结构相似性(structural similarity,SSIM)方面获得了更好的结果,进一步提高了超分辨率的细节特征。本文网络在获得相同的PSNR指标的情况下,模型参数减少了近50%。**结论**　通过极轴约束使得注意力对齐网络模型参数量大大减少,并能够充分捕获远距离信息来进行特征对齐,产生高效的时空特征,还通过设计动态融合机制,实现了高质量的重建结果。

关键词:视频超分辨率(VSR);轻量网络;可变形卷积;注意力约束;动态融合机制;残差空洞空间金字塔池化

Super-resolution video frame reconstruction through lightweight attention constraint alignment network

Jin Yutong, Song Huihui*, Liu Qingshan

Nanjing University of Information Science and Technology, Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing 210044, China

Abstract: **Objective** Current deep learning technology is beneficial to video super-resolution (SR) reconstruction. The existing methods are constrained of the accuracy of motion estimation and compensation based on optical flow estimation, and the reconstruction effect of large-scale moving targets is poor. The deformable convolutional alignment network captures the target's motion information via learning adaptive receptive fields, and provides a new solution for video super-resolution reconstruction. To reconstruct realistic high-resolution (HR) video frames, our lightweight-attention-constrained deformable alignment network aims to use a less model parameters network to make full use of the redundant information between the reference frame and adjacent frames. **Method** Our attention constraint alignment network (ACAN) consists of three key

收稿日期:2021-05-18;**修回日期**:2021-10-11;**预印本日期**:2021-10-18

*通信作者:宋慧慧　songhuihui@nuist.edu.cn

基金项目:国家自然科学基金项目(61872189,61532009);江苏省自然科学基金项目(BK20191397)

Supported by: National Natural Science Foundation of China (61872189, 61532009); Natural Science Foundation of Jiangsu Province, China (BK20191397)

components like feature extraction module, attention constraint alignment sub-network and dynamic fusion. First, the 5 layers are designed in terms of shared weights feature extraction module in the context of three ground residuals without batch normalization (BN) layer and two residuals atrous spatial pyramid pooling (res_ ASPP). To extract multi-scale information and multi-level information without increasing the amount of parameters, the two residuals atrous spatial pyramid pooling and three ground residuals are connected alternately without batch normalization layer. After that, the polar axis constraint and the attention mechanism are integrated to design a lightweight attention constraint alignment sub-network (ACAS). The network regulates the input features of deformable convolution via capturing the global correspondence between adjacent frames and reference frames in the time domain under polar axis constraints, and generates a reasonable offset to achieve implicit alignment. Specifically, the ACAS is introduced through combining the deformable convolution with attention and polar axis constraint. The three attention constraint blocks (ACB) involved ACAS to constrain the features on the horizontal axis of neighboring frames. To find out the most similar features, it can code the feature correlation between any two positions along the horizontal line. At the same time, an effective mask is designed to solve the unavoidable occlusion phenomenon in the video. In the feature extraction module, we send extracted features to the alignment module to generate alignment features with exact matching relationships. In the ablation experiment, we verified that the network can well capture the matching relationship between the reference frame and the adjacent frame using a layer of ACB. However, the network can capture the matching relationship between adjacent frames and the reference frame and handle the status of large motion in the video based on the cascaded three-layer ACB. Therefore, we select a cascaded three-layer ACB network during network design. We illustrate a dynamic fusion branch, which is composed of 16 dynamic fusion blocks. Each block is made of two spatial feature transformation (SFT) layers and two 1×1 convolutions. This branch fuses the time alignment features of the reference frame in the forward neural network and the spatial features of the original low-resolution (LR) frame at different stages. Finally, the high-resolution frame is reconstructed and to be trained. Vimeo-90K is a widely used training dataset and is evaluated in conjunction with the Vid4 test dataset in common. In the training process, this network is trained on Vimeo-90K dataset and tested on Vid4 and REDS4 datasets. The loss function chooses the Charbonnier penalty function solely. The channel size of each layer is set to 64 for the final comparison, where we designates that the alignment module is composed of a layer of attention constraint alignment module, while that the assigned alignment module is cascading from three layers of attention constraint alignment module. Additionally, the network makes use of seven consecutive frames as input. Our RGB patches of a size of 64×64 are used as input to the video SR, with the mini-batch size set to 16. We use the Adam optimizer to update the network parameters. The initial learning rate is set to $4e-4$. All experiments are conducted on PyTorch 1.0 and four Nvidia Tesla T4 GPUs. **Result** Our experiment is evaluated on two benchmark datasets quantitatively, including Vid4 and realistic and diverse scenes dataset (REDS4), and the proposed combined method obtained better results in the image quality indicators peak signal to noise ratio (PSNR) and structural similarity (SSIM). Our results are compared the model to 10 recognized super resolution models, including single image super resolution (SISR) and video super resolution (VSR) methods on two common datasets (Vid4, REDS4). The quantitative evaluations are involved of PSNR and SSIM, and the reconstructed images of each method are provided for comparison. Our reconstruction results show that the proposed model can recover precise details, and the effectiveness of the proposed alignment module with polar axis constraints is verified by comparing the results of no alignment operation and the results of one or three layers of attention constraint alignment. Without the use of alignment, the PSNR score is 22.11 dB, with one layer of ACB PSNR score increased by 1.81 dB, and with three layers of ACB, the PSNR score is increased by 1.21 dB. This result proves the effectiveness of attention constraint to aligning blocks, and the network of cascaded three-layer ACB can capture long-distance spatial information. The dynamic fusion (DF) module is also verified, and the comparative experiment shows that the DF module can improve the reconstruction performance. Our results demonstrate that the PSNR score on the Vid4 data set has increased by more than 0.33 dB compared to EDVR_M, which is an increase of about 1.2%. Compared with EDVR_M, the PSNR score has increased by 0.49 dB on the REDS4 dataset, which is an increase of about 1.6%. Moreover, under the condition of the same PSNR scores, the proposed model parameters are nearly 50% less than that of recurrent back-projection network (RBPN). Our PSNR value is much higher than dynamic upsampling filter (DUF) in terms of the same number of parameters. The PSNR is increased by 0.21 dB although the number of parameters is slightly higher than that of

EDVR_M in our model. **Conclusion** the number of model parameters is reduced dramatically in the attentional alignment network through the polar axis constraint. To achieve high quality reconstruction results, the distance information can be captured for feature alignment. It can integrate the spatio-temporal features of video frames.

Key words: video super resolution (VSR); lightweight network; deformable convolution; attentional constraint; dynamic fusion mechanism; residual atrous spatial pyramid pooling

0 引 言

视频超分辨率(video super-resolution, VSR)重建的目标是从LR(low-resolution)帧(参考帧)和其对应的多个相邻帧中恢复出逼真的HR(high-resolution)帧。视频超分重建应用十分广泛,例如视频监控、高清电视和视频后期制作等。Dai等人(2017)提出了可变形的卷积网络(deformable convolutional networks, DCNs),突破了卷积神经网络(convolutional neural networks, CNNs)中感受野采用固定几何结构的局限性。DCNs能够从目标任务中学习偏移量来增加空间采样位置,从而学习出自适应的感受野。随后,Zhu等人(2018)提出了DCNs的进阶版本DCNs v2,通过增强建模能力和更强的训练,提高其专注于相关图像区域的能力。随着DCNs的发展,其在视频超分重建领域取得重大突破。例如,Tian等人(2020)提出的TDAN(temporally-deformable alignment network)首次将DCNs应用到视频超分领域。TDAN网络无需计算光流,能够在特征层面自适应地对齐相邻帧。

传统的VSR算法通过考虑相邻LR帧之间的亚像素运动,将多个LR帧作为输入得到HR帧。Liu和Sun(2014)引入了贝叶斯方法,在重建原始的高分辨率帧的同时,估计底层运动、模糊核和噪声。Farsiu等人(2004)提出了一种基于双边先验知识来处理不同的数据和噪声模型。但是,由于这些方法是将输入的视频帧当做单幅图像进行重建,并没有考虑帧与帧之间的时序关系,极有可能无法处理连续帧。考虑到VSR的特性,对LR参考帧和相邻LR帧之间的时序关系进行建模对于提高重建性能至关重要。Tao等人(2017)提出了亚像素运动补偿(sub-pixel motion compensation, SPMC)层,并分析了该层在视频超分中的实用性,通过有效融合SPMC层与多帧信息来重建图像细节。Haris等人(2019)用反向投影网络(recurrent back-projection network, RBPN)从连续视频帧中整合时空上下文信息来精准对齐LR参考帧和相邻的LR帧。Wang等人(2019a)设计了一个带有可变形卷积的视频恢复框架(video restoration framework with enhanced deformable convolutions, EDVR),在特征级别上自适应地对齐参考帧和每个相邻帧,设计一个金字塔、级联和可变形(pyramid, cascading and deformable, PCD)对齐模块处理大尺度运动。以上方法尽管在重建性能方面获得大幅提升,但是还存在一些难题有待解决,其中,最主要的问题是特征对齐操作没有考虑帧间的长距离信息。若只采用扩大感受野的方式来获取长距离信息会导致GPU显存占用率高、网络模型过大的问题。如何设计一个参数量少的网络来捕获长距离信息成为一个亟待解决的问题。

为了解决上述问题,本文提出了一种基于轻量级注意力约束对齐网络的VSR方法,可在一定的先验条件约束下执行帧与帧之间的特征级别对齐操作,从而捕获长距离信息、减少计算力,且准确重建HR帧。具体地,受自注意力机制(Vaswani等,2017;Wang等,2018a)启发,本文网络将极轴约束与注意力机制结合,开发出一种轻量级注意力机制用来探索全局对应关系。对于参考帧中的每个像素,轻量级注意力机制会关注沿极轴方向的所有差异信息,并且学会聚焦于最相似的特征。实验结果表明,这种轻量级的注意力约束对齐网络的模型参数远小于对比方法,并在多个数据集上取得了优异性能。

本文的主要贡献总结如下:

1)提出一种轻量级注意力约束对齐网络,用于探索相邻帧与参考帧之间沿极轴方向的全局对应关系;

2)设计了一个多阶段的动态融合网络,用来融合前向神经网络中参考帧的时域对齐特征和原始LR帧在不同阶段的空间特征;

3)通过共享特征抽取层有效提取多层次信息,且在不增加参数量的情况下捕获视频帧中的多尺度信息;

4)本文算法在多个标准数据集上达到领先水平,并在相同的峰值信噪比(peak signal to noise ratio, PSNR)指标下,本文模型参数远小于对比方法。

1　注意力约束对齐网络

如图 1 所示,本文网络主要包含特征抽取模块、特征对齐子网络与动态融合分支 3 个部分。首先,对输入的视频帧序列进行双三次(bicubic,BI)下采样,将其和原来的视频帧序列一同输入特征抽取模块抽取特征。特征抽取模块的结构如图 1 中紫色虚线框所示。其中不带 BN(batch normalization)层的残差块结构残差空洞空间金字塔模块如图 1 所示。本文将不带 BN 层的残差块与残差空洞空间金字塔池化模块交替结合抽取多层次特征和多尺度特征。抽取的多层次特征充分利用高低语义间的互补优势,而抽取的多尺度特征用于捕获图像间的自相似性,特征抽取模块还通过共享网络权重降低计算量。接着,将特征抽取器捕获到的多帧信息输入到注意力约束对齐子网络(attention constraint alignment sub-network,ACAS)中,ACAS 结构如图 1 中蓝色虚线框所示,该网络通过捕获时域上相邻帧与参考帧在极轴约束下的全局对应关系来规范可变形卷积的输入特征,用以产生合理的偏移量来实现隐式对齐。随后,动态融合分支(dynamic fusion branch,DFB)(如图 1 中红色虚线框所示)在前向神经网络中将参考帧中的时序对齐特征和原始帧 $\boldsymbol{I}_t^{\mathrm{LR}}$ 在不同阶段的空间特征进行多级融合,最后通过双三次插值上采样重建出超分辨率结果 $\bar{\boldsymbol{I}}_t^{\mathrm{HR}}$ 。

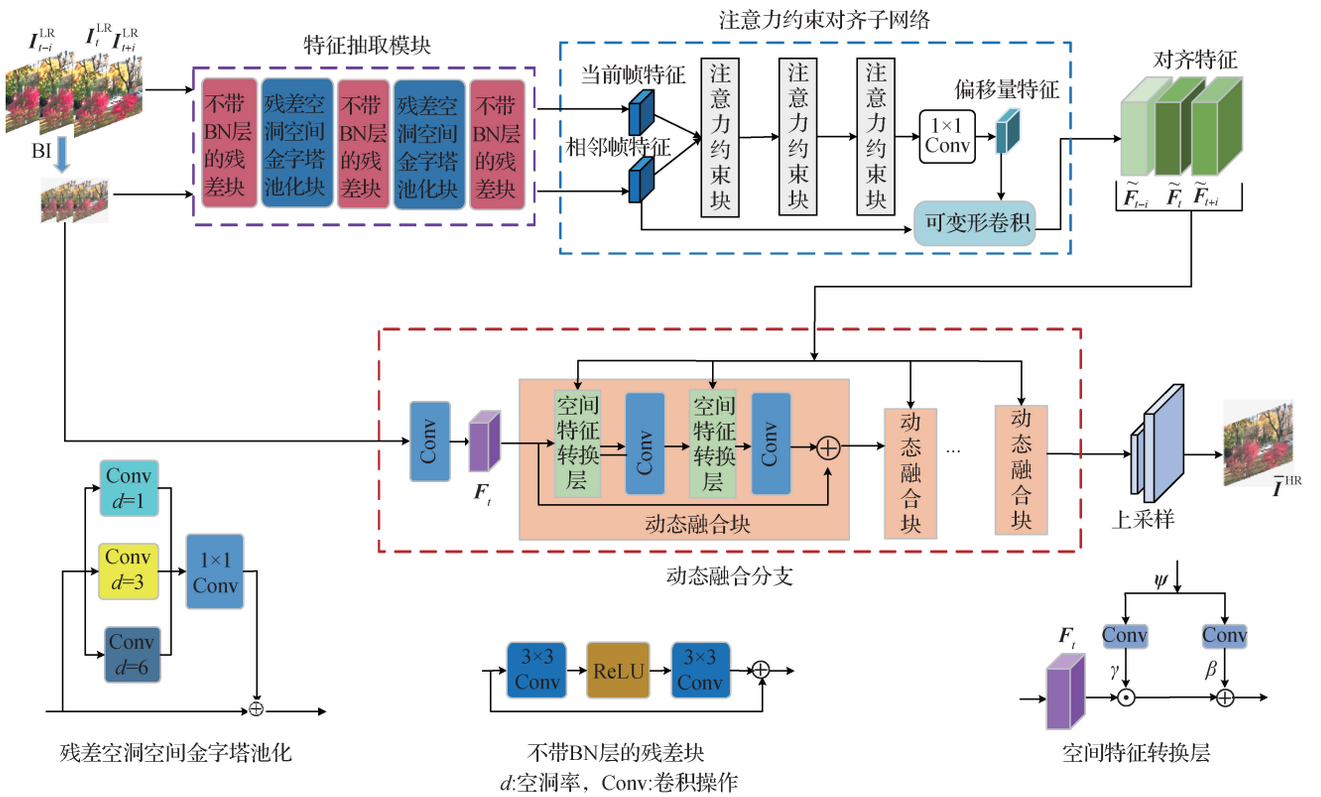


图 1　本文网络结构图

Fig. 1　Network structure diagram

本文的主要创新点在于所设计的注意力约束对齐子网络(ACAS)与动态融合分支(DFB)。注意力约束对齐子网络(ACAS)能够在极轴约束的条件下通过探索长距离信息捕获参考帧与相邻帧的全局对应关系,而动态融合分支(DFB)则能够动态地对时空特征进行融合。

1.1　注意力约束对齐子网络(ACAS)

受 Wang 等人(2019b)提出的视差注意力立体图像超分网络(parallax-attention stereo super resolution network,PASSRnet)和 Wang 等人(2022)提出的平行注意力机制(parallax-attention mechanism,PAM)的启发,本文提出了注意力约束块(attention

constraint block, ACB)来捕获相邻帧与参考帧之间的全局对应关系,用于生成合理的可变性卷积的偏移量。区别于 self-attention(Vaswani 等,2017)机制通过在特征图的横纵轴两个维度上变换来捕获全局对应关系,本文设计的注意力约束块通过极轴约束,只需要用一个维度的计算复杂度就可以捕获全局对应关系。

首先,特征抽取模块抽取到的特征为

$$[\boldsymbol{F}_{t-i},\boldsymbol{F}_t,\boldsymbol{F}_{t+i}] = E[\boldsymbol{I}_{t-i}^{\mathrm{LR}},\boldsymbol{I}_t^{\mathrm{LR}},\boldsymbol{I}_{t+i}^{\mathrm{LR}}] \tag{1}$$

式中,t 表示当前帧,又称为参考帧,$i \in (1,2,3,\cdots)$,E 表示特征抽取器,$\boldsymbol{I}_{t-i}^{\mathrm{LR}}$、$\boldsymbol{I}_t^{\mathrm{LR}}$、$\boldsymbol{I}_{t+i}^{\mathrm{LR}}$ 表示输入的低分辨率参考帧与其相邻帧,$\boldsymbol{F}_{t-i}$、$\boldsymbol{F}_t$、$\boldsymbol{F}_{t+i}$ 表示抽取到的参考帧和其相邻帧的特征。然后,将这些特征一同输入到注意力约束对齐子网络(ACAS)。其中的注意力约束块(ACB)如图 2 所示,具体地,给定一对图像的特征图 $\boldsymbol{M},\boldsymbol{N} \in \mathbf{R}^{H\times W\times C}$,其中,$\boldsymbol{M}$ 是相邻帧,$\boldsymbol{N}$ 是参考帧。$\boldsymbol{M}$ 被送到一个 1 × 1 的卷积层生成查询特征图 $\boldsymbol{K} \in \mathbf{R}^{H\times W\times C}$,与此同时,$\boldsymbol{N}$ 被送到另一个 1 × 1 卷积层得到 $\boldsymbol{J} \in \mathbf{R}^{H\times W\times C}$,将 $\boldsymbol{J}$ 维度变换为 $\mathbf{R}^{H\times C\times W}$,然后在 $\boldsymbol{K}$ 和 $\boldsymbol{J}$ 之间执行批量矩阵乘法,并应用到 softmax 层,得到注意力图 $\boldsymbol{L}_{N\to M} \in \mathbf{R}^{H\times W\times W}$,通过矩阵乘法注意力约束块可以有效地将沿极轴任意两个位置之间的特征相关性编码为注意力图。接下来,将 $\boldsymbol{N}$ 再送到 1 × 1 的卷积层生成响应特征映射 $\boldsymbol{H} \in \mathbf{R}^{H\times W\times C}$,$\boldsymbol{H}$ 和 $\boldsymbol{L}_{N\to M}$ 再做矩阵乘积产生输出特征 $\boldsymbol{S} \in \mathbf{R}^{H\times W\times C}$,$\boldsymbol{S}$ 作为所有可能差异特征的加权和。然后,将其与相应的局部特征 $\boldsymbol{M}$ 集成。值得注意的是,一旦得到 $\boldsymbol{L}_{N\to M}$,$\boldsymbol{M}$ 和 $\boldsymbol{N}$ 就交换生成 $\boldsymbol{L}_{M\to N}$,生成有效的掩码 $\boldsymbol{V}_N \in \mathbf{R}^{H\times W\times 1}$。由于注意力约束块(ACB)可以使用特征相似度逐步关注精确差异处的特征,因此可以捕获对应关系。在本文的网络设计中,为了保证网络轻量性的同时也能捕获远距离的对应关系,选择级联 3 层 ACB。在消融实验中,验证使用一层 ACB 的网络比没有进行对齐的网络在 PSNR 指标上获得了 1.81 dB 的增益,说明本文设计的 ACB 能够很好地对参考帧和相邻帧进行匹配。而使用级联 3 层 ACB 的网络比使用一层 ACB 的网络在 PSNR 指标上又提升了 1.21 dB。这证明级联 3 层 ACB 的网络能够更好地捕获相邻帧与参考帧之间的远距离对应关系,从而在性能上进一步提升。之后,将得到的对应关系和有效掩码送到一个 1 × 1 卷积层中进行特征融合生成合理的偏移量并且将其输入到可变形卷积中,对相邻帧 $\boldsymbol{M}$ 和偏移量进行匹配得到对齐特征

$$\widetilde{\boldsymbol{F}}_t = f_{\mathrm{ACAS}}(\boldsymbol{I}_{t-i}^{\mathrm{LR}},\boldsymbol{I}_t^{\mathrm{LR}},\boldsymbol{I}_{t+i}^{\mathrm{LR}}) \tag{2}$$

式中,$f_{\mathrm{ACAS}}(\cdot)$ 表示注意力约束对齐子网络提取对齐特征的操作。然后,将所有的相邻帧的对齐特征 $\widetilde{\boldsymbol{F}}_{t\pm N}$ 拼接起来得到时间对齐特征

$$\boldsymbol{\psi} = [\widetilde{\boldsymbol{F}}_{t-N},\cdots,\widetilde{\boldsymbol{F}}_{t-1},\widetilde{\boldsymbol{F}}_{t+1},\cdots,\widetilde{\boldsymbol{F}}_{t+N}] \tag{3}$$

为了获取可靠和一致的对应关系,本文引入了一致性来规范注意力约束对齐子网络(ACAS)。给定从一对图像 $\boldsymbol{M}$、$\boldsymbol{N}$ 中提取的特征表示,其中 $\boldsymbol{M}$ 表

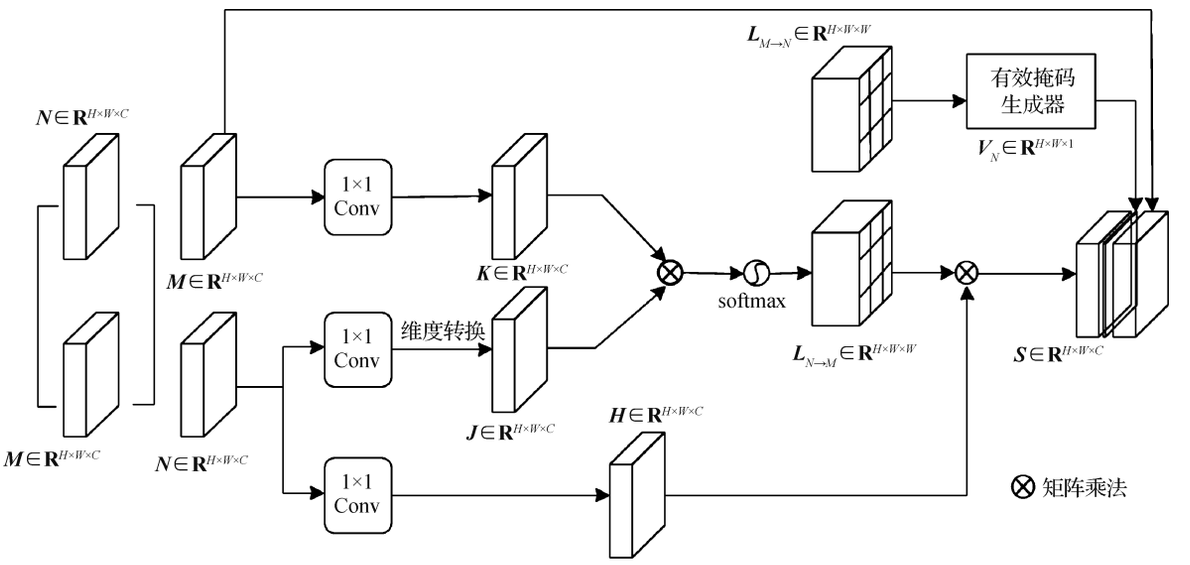


图 2　注意力约束块(ACB)

Fig. 2　Attention constraint block(ACB)

示相邻帧，$\boldsymbol{N}$ 表示参考帧，ACB 生成两个注意力图 $\boldsymbol{L}_{M\to N}$ 和 $\boldsymbol{L}_{N\to M}$。理想情况下，如果 ACB 捕获了准确的对应关系，则可以得到以下一致性

$$\begin{aligned}\boldsymbol{M} &= \boldsymbol{L}_{N\to M}\otimes\boldsymbol{N}\\ \boldsymbol{N} &= \boldsymbol{L}_{M\to N}\otimes\boldsymbol{M}\end{aligned}\tag{4}$$

式中，$\otimes$ 表示矩阵乘，$\boldsymbol{L}_{N(M)\to M(N)}$ 表示 $N(M)\to M(N)$ 的注意力图。另外，由于视频中不可避免地会出现遮挡现象，损害了一致性。为此，本文基于 $\boldsymbol{L}_{M\to N}$ 进行遮挡检测，生成有效掩码 $\boldsymbol{V}_N$，并且只对有效区域进行一致性正则化。在图 2 中，通常在注意力图中（如 $\boldsymbol{L}_{N\to M}$）为与遮挡区域相对应的垂直遮挡区域分配较小的权重。这是因为参考帧中的被遮挡像素与相邻帧的对应关系很少，因此，有效掩码 $V_N\in\mathbf{R}^{H\times W\times 1}$ 计算公式为

$$V_N(i,k) = \begin{cases}1 & \sum\limits_{j\in[1,W]} L_{N\to M}(i,j,k) > \tau\\ 0 & \text{其他}\end{cases}\tag{5}$$

式中，τ 为阈值（本文设置为 0.2）。

传统的注意力块（Wang 等，2018a）通过对 H 和 W 两个维度的变换来探索全局对应关系，不仅带来了巨大的参数量，而且 GPU 占用率高，不易训练。本文将极轴约束与注意力块相结合捕获极轴上的全局匹配关系，表 4 中的实验结果证明本文提出的极轴约束的注意力块能够带来很好的增益。而且它还大大降低了模型训练时的 GPU 内存占用，网络模型训练速度得到提升，最重要的是本文所设计的极轴约束对齐网络的参数量也比传统的注意力机制少。

1.2　动态融合分支（DFB）

简单的融合只发生在初始层，随着网络层数增加，来自相邻帧的互补时间信息将逐渐减弱（Kappeler 等，2016；Liao 等，2015）。受多阶段融合策略的启发（沈明玉 等，2019），本文提出一种动态融合方法解决上述问题，如图 1 底部分支所示。本文采用 Song 等人（2021）提出的调制特征融合模块中的一个子块拼接组成动态融合分支，并且参考 Wang 等人（2018b）提出的 SFTGAN（generative adversarial networks based on spatial feature transformation）网络确定本文的动态融合分支由 16 个共享权重的动态融合块组成。每个动态融合块如图 1 中淡橙色区域所示。它将式（3）中的时间对齐特征 $\boldsymbol{\psi}$ 作为共享条件来调制其输入参考帧的特征映射 $\boldsymbol{F}_t$。空间特征变换层（spatial feature transform，SFT）（Wang 等，2018b；Song 等，2021）结构见图 1，SFT 仿射变换为

$$f_{\mathrm{SFT}}(\boldsymbol{F}_t\mid\boldsymbol{\psi}) = \boldsymbol{\gamma}\odot\boldsymbol{F}_t+\boldsymbol{\beta}\tag{6}$$

式中，γ 和 β 分别表示尺度参数和平移参数，$\odot$ 表示像素级别的乘积，$f_{\mathrm{SFT}}(\cdot)$ 表示空间特征转换操作。γ 和 β 是由卷积层得到。在动态融合过程中，原始 LR 帧与对齐的时空特征同步增强参考帧的特征。最后，将高层次特征映射为高清图 $\hat{\boldsymbol{I}}_t^{\mathrm{HR}}$

$$\hat{\boldsymbol{I}}_t^{\mathrm{HR}} = f_{\mathrm{DFB}}(\boldsymbol{I}_t^{\mathrm{LR}}\mid\boldsymbol{\psi})\tag{7}$$

式中，$f_{\mathrm{DFB}}(\cdot)$ 表示动态融合操作。

2　实验设置和结果分析

2.1　实验设置

本文采用 Vimeo-90K（Xue 等，2019）作为训练集，Vid4（Liu 和 Sun，2014）和 REDS4（realistic and diverse scenes dataset）（Wang 等，2022）作为测试集训练和测试本文网络。损失函数为 $L=\sqrt{\|\hat{\boldsymbol{I}}_t^{\mathrm{HR}}-\boldsymbol{I}_t^{\mathrm{HR}}\|_2^2+\varepsilon^2}$（Lai 等，2019），其中 ε 设为 1E−3。网络以 7 个连续帧作为输入，裁剪为 64×64 像素的 RGB 补丁，小批量设为 16。本文使用 He 等人（2015）的方法初始化网络参数，并使用 $\beta_1=0.9$，$\beta_2=0.999$ 的 Adam 优化器（Kingma 和 Ba，2017）进行更新。初始学习速率设为 4E−4。Ours_S 和 Ours 分别表示对齐模块是由 1 层或 3 层 ACB 组成。所有实验在使用 PyTorch 1.0，4 张 NVIDIA Tesla T4 GPU 上进行。

2.2　结果分析

本文网络与 Bicubic、RCAN（residual channel attention networks）（Zhang 等，2018）和 DBPN（deep back-projection networks）（Haris 等，2018）、光流残差（吴昊 等，2021）、VESPCN（real-time video super-resolution with spatio-temporal networks and motion compensation）（Caballero 等，2017）、B_123 + T（Liu 等，2017）、SPMC（subpixel motion compensation networks）（Tao 等，2017）、TOFlow（task-oriented flow networks）（Xue 等，2019）、FRVSR（frame-recurrent video super-resolution）（Sajjadi 等，2018）、DUF（Jo 等，2018）、深度特征匹配（程松盛和潘金山，2021）、RBPN（Haris 等，2019）、EDVR（Wang 等，2022）进行比较。

表1显示了不同方法在4倍Vid4验证集的定量比较,包括PSNR和结构相似性(structural similarity, SSIM)(Wang等,2004)结果。Vid4是一个广泛使用的基准数据集,它包含4个视频序列:Calendar、City、Foliage和Walk,这些视频序列中包含有限的运动且高分辨率帧中存在伪影。由表1可以看出,本文网络的PSNR比EDVR_M方法高0.33 dB以上,约提升1.2%,且可以媲美RBPN网络。表2为REDS4数据集上所有方法的比较结果。REDS4是在NTIRE19挑战赛上发布的新的高质量数据集,由4个视频组成,分别为000、011、015、020,这些视频中包含更大更复杂的运动。由表2可以看出,本文方法获得最高的PSNR且比EDVR_M高出0.49 dB,约提升1.6%,PSNR和SSIM均与RBPN相当,更重要的是表3中显示的本文网络参数量远远小于RBPN。上述分析有力地证明本文方法可以通过探索长距离信息来捕获多帧之间的冗余特征,从而灵活地解决各种运动问题。

图3(a)演示了Vid4数据集中两个场景的可视化结果。从放大区域可以看出,本文网络重建出更

表1　不同方法在4倍Vid4验证集上的定量比较(PSNR/SSIM)

Table 1　Quantitative comparison of different methods in the 4 times Vid4 validation set (PSNR/SSIM)

算法	PSNR/(dB)/SSIM				平均值
	Calendar	City	Foliage	Walk	
Bicubic	20.39/0.572 0	25.16/0.602 8	23.47/0.566 6	26.10/0.797 4	23.78/0.634 7
DBPN	22.27/0.717 8	25.84/0.683 5	24.70/0.661 5	28.65/0.870 6	25.37/0.733 4
RCAN	22.33/0.725 4	26.10/0.696 0	24.74/0.664 7	28.65/0.871 9	25.46/0.739 5
VESPCN	–	–	–	–	25.35/0.755 7
B_123 + T	21.66/0.704 0	26.45/0.720 0	24.95/0.698 0	28.26/0.859 0	25.34/0.745 0
SPMC	22.16/0.746 5	27.00/0.7573	25.43/0.720 8	28.91/0.876 1	25.88/0.775 2
TOFlow	22.47/0.731 8	26.78/0.740 3	25.27/0.709 2	29.05/0.879 0	25.89/0.765 1
TDAN	22.98/0.756 0	26.99/0.757 0	25.51/0.717 0	29.50/0.890 0	26.24/0.780 0
光流残差	–	–	–	–	26.32/0.785 0
FRVSR	–	–	–	–	26.69/0.822 0
EDVR_M	23.36/0.790 9	27.57/0.794 0	25.99/0.747 3	30.23/0.902 4	26.79/0.808 7
Ours_S	23.79/0.800 9	27.61/0.797 0	26.07/0.751 6	30.52/0.907 8	27.00/0.814 3
RBPN	**23.93**/0.803 0	27.64/**0.802 0**	**26.27**/**0.757 0**	30.65/**0.911 0**	**27.12**/0.818 0
Ours	23.92/**0.806 7**	**27.69**/0.801 3	26.15/0.755 6	**30.70**/0.910 5	**27.12**/**0.818 5**

注:加粗字体为每列最优结果,下划线字体为每列次优结果,"–"表示该方法未测各指标。

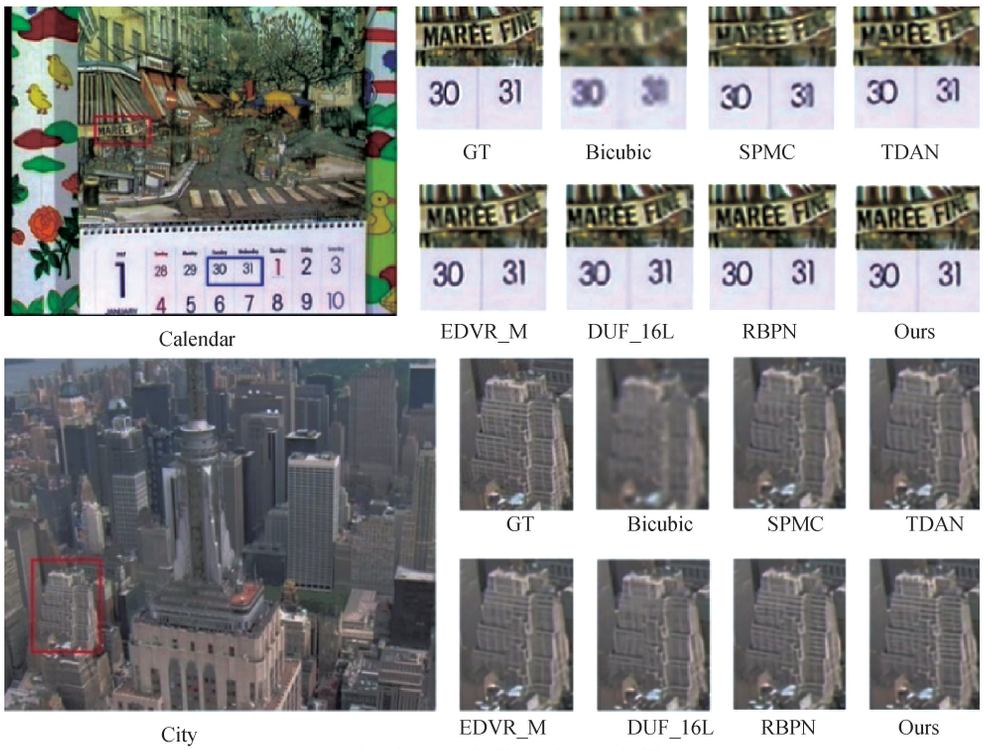
表2　不同方法在4倍REDS4测试集的评估结果(PSNR/SSIM)

Table 2　Evaluation results of different methods in 4 times REDS4 test set (PSNR/SSIM)

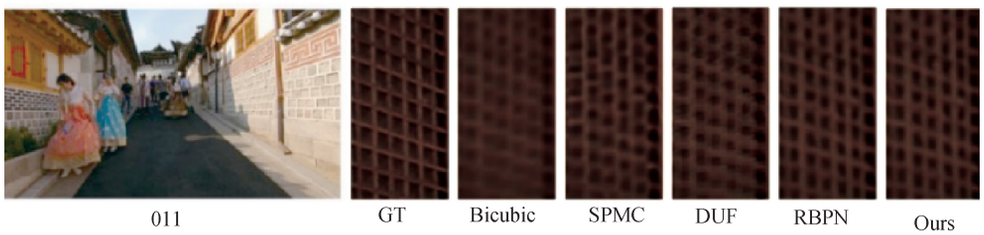
视频	PSNR/(dB)/SSIM							
	Bicubic	RCAN	SPMC	DUF	EDVR_M	深度特征匹配	RBPN	本文
000	20.55/0.531 2	21.89/0.543 6	22.16/0.586 4	22.58/0.597 1	28.12/0.819 5	27.85/0.811 2	**28.86**/0.819 5	**28.86**/**0.819 6**
011	21.68/0.653 5	22.67/0.660 2	23.11/0.667 9	23.32/0.680 7	30.45/**0.861 9**	31.49/0.875 6	31.08/0.859 2	**31.09**/0.859 3
015	24.17/0.545 7	24.83/0.590 1	25.05/0.613 4	25.71/0.776 4	33.81/0.919 9	33.47/0.912 5	34.18/0.919 9	**34.19**/**0.920 0**
020	21.21/0.668 7	21.94/0.671 5	22.13/0.675 3	22.91/0.687 2	29.56/**0.865 6**	**29.80**/0.878 1	29.78/0.864 2	29.78/0.864 2
平均值	21.90/0.599 7	22.83/0.616 3	23.11/0.635 7	23.63/0.685 4	30.49/0.867 2	30.55/**0.869 3**	30.97/0.865 7	**30.98**/0.865 8

注:加粗字体为每行最优结果,下划线字体为每列次优结果。

精细、更可靠的细节。在 Calendar 视频的帧示例中，恢复出最清晰的数字 31。在 City 视频的帧示例中，本文方法与 RBPN 均能对密集的大楼纹理外观进行重建。图 3(b)展示了 REDS4 数据集上的可视化结果，可以看出本文方法能够较清晰地区分出窗户部分的细节，RBPN 虽然也能较为清晰地重建出这些细节特征，但其网络结构较本文网络而言更为复杂。由表 3 可知，本文网络参数量仅为其二分之一。以上分析充分证明本文框架能够在大大减少计算量的情况下大幅提升视觉质量。



(a) 在 4 倍 Vid4 数据集上的定性结果



(b) 在 4 倍 REDS4 数据集上的定性结果

图 3　可视化结果

Fig. 3　Visual of result ((a) super-resolution reconstruction results on the 4 times Vid4 dataset; (b) super-resolution reconstruction results on the 4 times REDS4 dataset)

2.3　模型大小的比较

表 3 显示了本文方法与 DBPN、RCAN、EDVR_M、DUF、RBPN 的参数对比情况。DBPN 和 RCAN 是目前两种最好的 SISR 方法，但它们都有较大的模型尺寸，参数量达 1 000 多万。表 3 表明 RBPN 参数量在 VSR 方法中是最多的。结合表 2 中的平均值来看，在 PSNR 值相当的情况下，模型 Ours 的参数量比 RBPN 少了近 50%。在参数量相当的情况下，模型 Ours 的 PSNR 值远远高于 DUF。而模型 Ours_S 的参数量虽然略高于 EDVR_M，但是 PSNR 提高了 0.21 dB(见表 1)。这证明本文网络在参数量小的情况下取得了优异的性能，实现了轻量级的

表 3　不同方法的模型参数比较

Table 3　Comparison of model parameters of different methods

	DBPN	RCAN	EDVR_M	DUF	RBPN	Ours_S	Ours
参数量/M	10	16	4.2	5.9	12.5	4.9	5.8

网络设计。

2.4 消融实验

本文对注意力约束对齐模块和动态融合模块进行验证。消融实验结果在 Vid4 数据集上测得。首先,将 ACB 移除并替换为简单的卷积操作,称之为 Baseline。表 4 表明在 Baseline 获得最低的 PSNR 值,在 Baseline 中加入一层 ACB,模型 ACB-1 的 PSNR 指标提高到 23.92 dB,增益为 1.81 dB。而将 ACB 级联 3 层加入 Baseline 中,模型 ACB-3 的 PSNR 指标达到 25.13 dB,比 ACB-1 提高了 1.21 dB。这解释了 3 层注意力约束对齐模块能够很好地捕获大运动,即能够比 ACB-1 更好地捕获远距离对应关系。另外,为了验证动态融合模块的有效性,在 ACB-3 模型后面接入 16 层动态融合块,模型 DF 的 PSNR 指标达到 26.28 dB,增益为 1.35 dB,这证明在特征融合过程的每个阶段逐步增强参考帧的特征,可以实现更准确的重建结果。

表 4 消融实验

Table 4 Ablation experiments

Baseline	ACB-1	ACB-3	DF	PSNR/dB
√	–	–	–	22.11
√	√	–	–	23.92
√	–	√	–	25.13
√	–	√	√	26.48

注:“√”表示采用,“–”表示未采用。

3 结 论

本文提出了一种轻量级注意力约束对齐网络的视频超分重建算法,在大量减少模型参数量的同时又能高效且准确地进行超分重建,文中对比实验证明了其有效性和优越性。本文的创新点总结如下:1)通过一个共享权重的特征提取器提取输入帧中丰富的多层次信息。2)在极轴约束的前提条件下,设计一个轻量的注意力对齐块使网络能够关注特征图水平轴上所有特征中最相似的特征,实现精准对齐。针对存在大运动的视频,设计一个级联 3 层注意力约束块的网络捕获远距离信息以生成规范的偏移量,将其与相邻帧送入可变形卷积中实现精准对齐。3)用 16 层共享权重的动态融合块组成的动态融合分支充分融合相邻帧的时间对齐特征和原始 LR 帧在不同阶段的空间特征。最后上采样重建出高分辨率视频帧。实验表明,本文方法在两个基准测试数据集上超过了先进的视频超分算法,能够提升视频帧的超分辨率细节特征,并且大大减少了参数量。

然而,由于现有的视频超分算法的数据集有限,大部分模型旨在找到现有数据集的特性以此获得较好的结果,而在真实场景中往往存在多种不确定情况,比如未知的噪声、模糊等,如何应对这些未知情况关乎着超分算法能否落地,因此本文将进一步研究真实场景的超分,针对真实情况中的模糊噪声叠加的问题设计解决方案,继续改善算法性能。

参考文献(References)

Caballero J, Ledig C, Aitken A, Acosta A, Totz J, Wang Z H and Shi W Z. 2017. Real-time video super-resolution with spatio-temporal networks and motion compensation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 2848-2857 [DOI: 10.1109/CVPR.2017.304]

Cheng S S and Pan J S. 2021. Video super-resolution method based on deep learning feature warping. Computer Science, 48(7): 184-189 (程松盛, 潘金山. 2021. 基于深度学习特征匹配的视频超分辨率方法. 计算机科学, 48(7): 184-189)

Dai J F, Qi H Z, Xiong Y W, Yi L, Zhang G D, Hu H and Wei Y C. 2017. Deformable convolutional networks//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 764-773 [DOI: 10.1109/ICCV.2017.89]

Farsiu S, Robinson M D, Elad M and Milanfar P. 2004. Fast and robust multiframe super resolution. IEEE Transactions on Image Processing, 13(10): 1327-1344 [DOI: 10.1109/TIP.2004.834669]

Haris M, Shakhnarovich G and Ukita N. 2018. Deep back-projection networks for super-resolution//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1664-1673 [DOI: 10.1109/CVPR.2018.00179]

Haris M, Shakhnarovich G and Ukita N. 2019. Recurrent back-projection network for video super-resolution//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 3892-3901 [DOI: 10.1109/CVPR.2019.00402]

He K M, Zhang X Y, Ren S Q and Sun J. 2015. Delving deep into rectifiers: surpassing human-level performance on imagenet classification//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE: 1026-1034 [DOI: 10.1109/ICCV.2015.123]

Jo Y, Oh S W, Kang J and Kim S J. 2018. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE:

3224-3232［DOI：10.1109/CVPR.2018.00340］

Kappeler A，Yoo S，Dai Q Q and Katsaggelos A K. 2016. Video super-resolution with convolutional neural networks. IEEE Transactions on Computational Imaging，2（2）：109-122［DOI：10.1109/TCI.2016.2532323］

Kingma D P and Ba J. 2017. Adam：a method for stochastic optimization［EB/OL］.［2021-05-03］. https://arxiv.org/pdf/1412.6980.pdf

Lai W S，Huang J B，Ahuja N and Yang M H. 2019. Fast and accurate image super-resolution with deep laplacian pyramid networks. IEEE Transactions on Pattern Analysis and Machine Intelligence，41（11）：2599-2613［DOI：10.1109/TPAMI.2018.2865304］

Liao R J，Tao X，Li R Y，Ma Z Y and Jia J Y. 2015. Video super-resolution via deep draft-ensemble learning//Proceedings of 2015 IEEE International Conference on Computer Vision（ICCV）. Santiago，Chile：IEEE：531-539［DOI：10.1109/ICCV.2015.68］

Liu C and Sun D Q. 2014. On Bayesian adaptive video super resolution. IEEE Transactions on Pattern Analysis and Machine Intelligence，36（2）：346-360［DOI：10.1109/TPAMI.2013.127］

Liu D，Wang Z W，Fan Y C，Liu X M，Wang Z Y，Chang S Y and Huang T. 2017. Robust video super-resolution with learned temporal dynamics//Proceedings of 2017 IEEE International Conference on Computer Vision（ICCV）. Venice，Italy：IEEE：2526-2534［DOI：10.1109/ICCV.2017.274］

Sajjadi M S M，Vemulapalli R and Brown M. 2018. Frame-recurrent video super-resolution//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City，USA：IEEE：6626-6634［DOI：10.1109/CVPR.2018.00693］

Shen M Y，Yu P F，Wang R G，Yang J and Xue L X. 2019. Image super-resolution reconstruction via deep network based on multi-staged fusion. Journal of Image and Graphics，24（8）：1258-1269（沈明玉，俞鹏飞，汪荣贵，杨娟，薛丽霞. 2019. 多阶段融合网络的图像超分辨率重建. 中国图象图形学报，24（8）：1258-1269）［DOI：10.11834/jig.180619］

Song H H，Xu W J，Liu D，Liu B，Liu Q S and Metaxas D N. 2021. Multi-stage feature fusion network for video super-resolution. IEEE Transactions on Image Processing，30：2923-2934［DOI：10.1109/TIP.2021.3056868］

Tao X，Gao H Y，Liao R J，Wang J and Jia J Y. 2017. Detail-revealing deep video super-resolution//Proceedings of 2017 IEEE International Conference on Computer Vision（ICCV）. Venice，Italy：IEEE：4482-4490［DOI：10.1109/ICCV.2017.479］

Tian Y P，Zhang Y L，Fu Y and Xu C L. 2020. TDAN：temporally-deformable alignment network for video super-resolution//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE：3357-3366［DOI：10.1109/CVPR42600.2020.00342］

Vaswani A，Shazeer N，Parmar N，Uszkoreit J，Jones L，Gomez A N，Kaiser L and Polosukhin I. 2017. Attention is all you need［EB/OL］.［2021-05-12］. https://arxiv.org/pdf/1706.03762.pdf

Wang L G，Guo Y L，Wang Y Q，Liang Z F，Lin Z P，Yang J G and An W. 2022. Parallax attention for unsupervised stereo correspondence learning. IEEE Transactions on Pattern Analysis and Machine Intelligence，44（4）：2108-2125［DOI：10.1109/TPAMI.2020.3026899］

Wang L G，Wang Y Q，Liang Z F，Lin Z P，Yang J G，An W and Guo Y L. 2019b. Learning parallax attention for stereo image super-resolution//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Long Beach，USA：IEEE：12242-12251［DOI：10.1109/CVPR.2019.01253］

Wang X L，Girshick R，Gupta A and He K M. 2018a. Non-local neural networks［EB/OL］.［2021-05-12］. https://arxiv.org/pdf/1711.07971.pdf

Wang X T，Chan K C K，Yu K，Dong C and Loy C C. 2019a. EDVR：video restoration with enhanced deformable convolutional networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops（CVPRW）. Long Beach，USA：IEEE：1954-1963［DOI：10.1109/CVPRW.2019.00247］

Wang X T，Yu K，Dong C and Loy C C. 2018b. Recovering realistic texture in image super-resolution by deep spatial feature transform//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City，USA：IEEE：606-615［DOI：10.1109/CVPR.2018.00070］

Wang Z，Bovik A C，Sheikh H R and Simoncelli E P. 2004. Image quality assessment：from error visibility to structural similarity. IEEE Transactions on Image Processing，13（4）：600-612［DOI：10.1109/TIP.2003.819861］

Wu H，Lai H C，Qian X Z and Chen H. 2021. Video super－resolution reconstruction algorithm based on optical flow residuals. Computer Engineering and Applications，58（15）：220-228（吴昊，赖惠成，钱绪泽，陈豪. 2021. 基于光流残差的视频超分辨率重建算法. 计算机工程与应用，58（15）：220-228）［DOI：10.3778/j.issn.1002-8331.2012-0409］

Xue T F，Chen B A，Wu J J，Wei D L and Freeman W T. 2019. Video enhancement with task-oriented flow. International Journal of Computer Vision，127（8）：1106-1125［DOI：10.1007/s11263-018-01144-2］

Zhang Y L，Li K P，Li K，Wang L C，Zhong B N and Fu Y. 2018. Image super-resolution using very deep residual channel attention networks//Proceedings of the 15th European Conference on Computer Vision. Munich，Germany：Springer：294-310［DOI：10.1007/978-3-030-01234-2_18］

Zhu X Z，Hu H，Lin S and Dai J F. 2018. Deformable ConvNets v2：more deformable，better results//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach，USA：IEEE：9300-9308［DOI：10.1109/CVPR.2019.00953］

作者简介

靳雨桐，女，硕士研究生，主要研究方向为超分辨率重建、深度学习。E-mail：1451185284@qq.com

宋慧慧，通信作者，女，教授，博士生导师，主要研究方向为视频目标分割、图像超分。E-mail：songhuihui@nuist.edu.cn

刘青山，男，教授，博士生导师，主要研究方向为视频内容分析与理解。E-mail：qsliu@nuist.edu.cn