计算机应用
Journal of Computer Applications
ISSN 1001-9081,CN 51-1307/TP

《计算机应用》网络首发论文

题目：基于注意力融合网络的视频超分辨率重建
作者：卞鹏程，郑忠龙，李明禄，何依然，王天翔，张大伟，陈丽媛
收稿日期：2020-08-24
网络首发日期：2020-10-22
引用格式：卞鹏程，郑忠龙，李明禄，何依然，王天翔，张大伟，陈丽媛．基于注意力融合网络的视频超分辨率重建[J/OL]．计算机应用．
https://kns.cnki.net/kcms/detail/51.1307.tp.20201021.0852.004.html

Journal of Computer Applications　ISSN 1001-9081
计算机应用　CODEN JYIIDU　http://www.joca.cn

DOI:10.11772/j.issn.1001-9081.2020081292

基于注意力融合网络的视频超分辨率重建

卞鹏程,郑忠龙*,李明禄,何依然,王天翔,张大伟,陈丽媛
(浙江师范大学 数学与计算机科学学院,浙江 金华 321004)
(*通信作者电子邮箱 zhonglong@zjnu.edu.cn)

摘　要:基于深度学习的视频超分辨率方法主要关注视频特征的帧内和帧间时空关系,但以往的方法在视频帧的特征对齐和融合方面存在运动信息估计不精确、特征融合不充分等问题。针对这些问题,采用反向投影原理,结合多种注意力机制和融合策略,构建了一个基于注意力融合网络(AFN)的视频超分辨率模型。首先,在特征提取阶段,为了处理相邻帧和参考帧之间的多种运动,采用反向投影结构获取运动信息的误差反馈;然后,使用时间、空间和通道注意力融合模块用于多维度的特征挖掘和融合;最后,在重建阶段,将得到的高维特征经过卷积,重建出高分辨率的视频帧。通过学习视频帧内和帧间特征的不同权重,充分挖掘视频帧的相关关系,并利用迭代网络结构,采取渐进的方式由粗到精地处理提取到的特征。在两个公开的基准数据集上的实验结果表明,注意力融合网络能够有效处理包含多种运动和遮挡的视频,与一些主流方法相比在量化指标上提升较大,如对于4倍重建任务,AFN产生的视频帧的峰值信噪比(PSNR)指标在Vid4数据集上比帧循环视频超分辨率网络(FRVSR)提高了13.2%,在SPMCS数据集上比动态上采样滤波视频超分辨率网络(VSR-DUF)模型提高了15.3%。

关键词:超分辨率;注意力机制;特征融合;反向投影;视频重建
中图分类号:TP391.4　　**文献标志码:**A

Attention fusion network based video super-resolution reconstruction

BIAN Pengcheng, ZHENG Zhonglong*, LI Minglu, HE Yiran, WANG Tianxiang, ZHANG Dawei, CHEN Liyuan
(*College of Mathematics and Computer Science, Zhejiang Normal University, Jinhua Zhejiang* 321004, *China*)

Abstract: Video super-resolution methods based on deep learning mainly focus on the spatial and temporal relationships of inter-frame and intra-frame in the video, but previous methods have many shortcomings in the feature alignment and fusion of video frames, such as inaccurate motion information estimation and insufficient feature fusion. Aiming at these problems, an Attention Fusion Network (AFN) was proposed for video super-resolution, making use of the back-projection principle combined with multiple attention mechanisms and fusion strategies. Specifically, in the feature extraction stage, in order to deal with various motion between neighbor frames and reference frame, a back-projection architecture was exploited to obtain the error feedback of motion information, then a temporal, spatial and channel attention fusion module was used to explore and fuse informative features in different dimensions. Finally, the high-dimensional features were convoluted to reconstruct high-resolution video frames. By learning different weights of features within and between video frames, the correlation between video frames was fully explored, and an iterative network structure was adopted to process the extracted features gradually from coarse to fine. The experimental results on two publicly available benchmark datasets show that the attention fusion network can effectively process videos that contain a variety of complex motions and occlusions, and achieves significant improvements in quantitative indicators over some mainstream methods, for instance, for video reconstruction with the scale factor set to 4, the Peak Signal to Noise Ratio (PSNR) of the frame reconstructed by AFN is 13.2% higher than that of Frame Recurrent Video Super-Resolution network (FRVSR) on Vid4 dataset and 15.3% higher than Video Super-Resolution network using Dynamic Upsampling Filter (VSR-DUF) on SPMCS dataset.

Key words: super-resolution; attention mechanism; feature fusion; back-projection; video reconstruction

0　引言

超分辨率(Super-Resolution, SR)技术的主要目的是通过对低分辨率(Low-Resolution, LR)图像填充丢失的细节信息重建出对应的高分辨率(High-Resolution, HR)图像。超分辨率技术可以分为单幅图像超分辨率(Single Image Super-

收稿日期:2020-08-24;**修回日期:**2020-09-18;**录用日期:**2020-10-13。　**基金项目:**国家自然科学基金资助项目(61672467)。
作者简介:卞鹏程(1993—),男,安徽六安人,硕士研究生,主要研究方向:深度学习、计算机视觉;　郑忠龙(1976—),男,河北沧州人,教授,博士,CCF会员,主要研究方向:模式识别、机器学习、图像处理;李明禄(1965—),男,重庆人,教授,博士,CCF会员,主要研究方向:云计算、车辆自组网络、无线传感器网络、大数据分析;　何依然(1996—),女,浙江杭州人,硕士,主要研究方向:机器学习;王天翔(1994—),男,浙江金华人,博士研究生,主要研究方向:机器学习、计算机视觉;　张大伟(1995—),男,江苏宿迁人,博士研究生,主要研究方向:深度学习、计算机视觉;　陈丽媛(1994—),女,河南焦作人,博士研究生,主要研究方向:深度学习、计算机视觉。

Resolution，SISR）[1-6]重建、多幅图像超分辨率（Multi-Image Super-Resolution，MISR）[7-8]重建和视频超分辨率（Video Super-Resolution，VSR）[9-12]重建。SISR重建技术主要是通过利用图像先验或者图像内部的自相似性进行重建，而VSR重建技术不仅可以利用帧内的空间信息，也可以通过显式或隐式的特征对齐进行运动估计和运动补偿，挖掘相邻帧与参考帧之间的时间关系，用于指导目标帧的重建。超分辨率技术应用广泛，主要应用领域有医学影像处理、视频监控、遥感卫星图像处理、刑事案件侦破和超高清产业[13]。

由于超分辨率重建是一种不适定问题，对于单一的低分辨率图像或者视频帧，可能存在许多不同的高分辨率图像与之对应。利用深度学习算法，模型可以直接学习从低分辨率图像到高分辨率图像的端到端映射函数，从而对生成高分辨率图像的过程进行指导和约束。目前，由于深度学习的不断发展，出现了很多基于深度学习的超分辨率重建方法，研究人员在特征提取、非线性映射和重建的架构，以及损失函数、学习策略和评价指标等各个方面开展了广泛的研究，也取得了不断的突破[14]。

现有的基于深度学习的视频超分辨率方法一般有特征提取、对齐、融合和重建四个步骤，特征对齐和融合主要是对多帧进行显式或隐式的运动估计和运动补偿，当视频中存在遮挡、复杂的运动等问题时，特征对齐和融合的策略对视频帧重建的质量起着关键的作用[12]。由于相邻的视频帧存在大量的冗余信息，为了充分挖掘视频帧内的空间关系以及帧间的时间关系来达到更好的特征融合，不仅需要深度网络能够尽可能地增大参考帧特征的感受野来提取不同层的空间特征，也需要有选择地利用不同的相邻帧所提供的不同信息量。以往的方法在特征对齐和融合方面存在很多不足。例如，在使用光流进行运动估计和运动补偿时，往往很难得到精确的光流，不精确的运动估计会严重影响后续的超分辨率重建效果；在进行特征融合时，没能充分利用特征间的关系。

针对以上的问题，本文以反向投影架构作为骨干网络，迭代地学习特征之间的映射关系，并使用多种注意力机制对帧内信息和帧间信息关系进行利用，构建了一个统一的基于注意力机制的超分辨率网络。通过学习和融合帧间信息和帧内信息，聚合提取到的特征来增强模型的表征能力。值得注意的是，本文提出的注意力模块也可以应用到其他的视频超分辨率网络结构中。最后，在两个公开的数据集上与一些主流的视频超分辨率方法进行对比实验，结果表明本文提出的模型在视觉质量和量化指标上都有很好的竞争力。同时，相关的消融实验也证明了注意力融合模块的有效性。图1(b)展示了本文的注意力融合网络（Attention Fusion Network，AFN）对双三次插值下采样后的LR视频帧进行4倍超分辨率重建后的结果。

(a) Bicubic (b) AFN (c) HR

图1 双三次插值、AFN和高分辨率原图4倍放大对比

Fig. 1 Comparison of the bicubic interpolation, AFN and the HR with the scale ratio set to 4.

1 相关工作

1.1 基于深度学习的视频超分辨率

随着深度学习的发展，深度学习方法在计算机视觉任务中的应用也越来越广泛。图像和视频超分辨率重建是一种底层计算机视觉任务，较早基于深度学习的超分辨率方法是2014年Dong等[15]提出的超分辨率卷积神经网络（Super-Resolution Convolutional Neural Network，SRCNN）。相较于图像超分辨率，视频超分辨率的研究更加关注帧间显式或隐式的对齐和多帧融合，Kappeler等[16]首次使用光流进行运动补偿，然后将多帧串联送入卷积网络。Caballero[17]提出的视频高效亚像素卷积神经网络（Video Efficient Sub-Pixel Convolutional neural Network，VESPCN）也使用光流进行运动估计和运动补偿，并使用亚像素卷积进行上采样，Tao等[18]提出了亚像素运动补偿模块，并使用编码解码结构结合卷积长短期记忆网络（Convolutional Long-Short Term Memory network，ConvLSTM）[19]加速训练和细节融合。Liu等[20]提出时间自适应网络选取最优范围的时间依赖来处理不同运动。Sajjadi等[11]提出了帧循环视频超分辨率网络（Frame Recurrent Video Super-Resolution network，FRVSR）去逐步融合多帧，以确保时间连续性。Younghyun等[21]提出的动态上采样视频超分辨率网络（Video Super-Resolution network using Dynamic Upsampling Filter，VSR-DUF）构建了一个三维卷积[22]模块，并采用动态上采样滤波器对目标帧进行重建。Haris等[23]提出的循环反向投影网络（Recurrent Back-Projection Network，RBPN）采用循环编码解码结构，利用反向投影机制估计隐式的帧间运动以及LR视频帧和HR视频帧之间的映射关系。Yi等[24]提出的非局部时空关系渐进融合网络（Progressive Fusion video super-resolution network via exploiting Non-Local spatio-temporal correlations，PFNL）采用非局部注意力机制（Non-local Attention）[25]挖掘时空关系，并采用渐进的方式逐步融合特征。Wang等[12]提出的增强可变卷积视频恢复网络（Video Restoration framework with Enhanced Deformable convolutions，EDVR）采用可变卷积进行特征对齐，结合金字塔结构进行时间和空间信息融合。传统的采用光流作为特征对齐的方法不仅费时而且不精确的光流估计会影响后续的特征融合和重建。本文采用反向投影原理，对特征进行隐式对齐，并结合多种注意力机制进行特征融合，以克服传统光流法的不足。

1.2 注意力机制

注意力机制模仿人类视觉机制，通过选取局部重点关注区域进而提取主要特征，忽略次要特征。近几年，注意力机制在计算机视觉和自然语言处理等领域都取得了重要的突破。通过在多种视觉任务中嵌入注意力机制，可以提升深度学习模型的表现性能。根据注意力关注的域，可以分为空间域、通道域、时间域和混合域。Jaderberg等[26]提出的空间变换网络（Spatial Transform Networks，STN）使用空间域注意力将原始图像的空间信息变换到另一个空间并保留了关键信息。Hu等[27]在挤压激励网络（Squeeze and Excitation Network，SENet）中提出通道注意力，研究特征通道之间的关系，对通道维度的信息进行利用。Wang等[28]提出的高效通道注意力网络（Efficient Channel Attention Network，ECA-Net）通过改进SENet，在提升性能的同时，更加轻量化。Wang等[29]提出的残

差注意力网络通过上采样、下采样和残差模块,形成了空间域和通道域的混合注意力。非局部注意力网络[25]通过捕获长范围的特征依赖性,分析了当前像素与全局的其他像素之间关系权重。在超分辨率任务中引入注意力机制的研究也很多,Liu等[30]提出使用注意力机制区分图像的纹理与平滑区域。Zhang等[31]提出的残差通道注意力网络(Residual Channel Attention Network, RCAN)使用残差注意力来提高模型的表征能力。Liu等[32]提出的基于注意力的反投影网络(Attention based Back-Projection Network, ABPN)采用非局部注意力来获取图像内部像素之间的空间关系。Liu等[33]提出残差特征聚合网络结合增强空间注意力来提高图像超分辨率重建的效果。EDVR采用时间空间注意力来进行特征融合,并采用金字塔结构增加注意力感受野。本文提出采用时间、空间和通道注意力对特征的相关关系进行更充分的挖掘和利用,提高模型的特征处理能力。

2 注意力融合网络

2.1 概述

给定$2n+1$个连续的低分辨率视频帧,表示为$\{\boldsymbol{I}_{t-n}^{\mathrm{lr}}:\boldsymbol{I}_{t+n}^{\mathrm{lr}}\}$,每帧的大小为$M\times N$,其中$M$表示视频帧的高度,$N$表示宽度。设定参考帧为$\boldsymbol{I}_t^{\mathrm{lr}}$,其余帧称为$\boldsymbol{I}_t^{\mathrm{lr}}$的邻帧,视频超分辨率的目标是根据参考帧和邻帧,重建出与参考帧对应的高分辨率视频帧$\hat{\boldsymbol{I}}_t^{\mathrm{sr}}$,大小为$sM\times sN$,其中$s$为比例因子,要求预测的高分辨率视频帧$\hat{\boldsymbol{I}}_t^{\mathrm{sr}}$与相应的实际高分辨率视频帧$\boldsymbol{I}_t^{\mathrm{hr}}$尽可能接近。低分辨率视频帧是由高分辨率视频帧下采样降质得到的,一般采用双三次插值下采样。图像退化模型的数学表达式为:

$$\boldsymbol{I}^{\mathrm{lr}}=\mathrm{SB}\boldsymbol{I}^{\mathrm{hr}}+u=\mathrm{D}\boldsymbol{I}^{\mathrm{hr}}+u \tag{1}$$

其中:$\boldsymbol{I}^{\mathrm{lr}}$为退化后的图像,$\boldsymbol{I}^{\mathrm{hr}}$为原始图像,S表示下采样操作,B表示模糊操作,$u$表示噪声,D表示将下采样和模糊操作统一表示为图像降质的过程。

本文利用反向投影原理,结合多种注意力机制和融合策略,构建了一个统一的用于视频超分辨率重建的注意力融合网络(AFN),AFN整体网络框架如图2所示,模型主要包括反向投影模块、注意力融合模块和重建模块,分别用于特征的提取、融合和重建。

具体的反向投影模块和注意力融合模块的结构将在2.2节、2.3节进行详细介绍。通过两个反向投影模块,可以迭代地最小化重建损失,学习高频特征和低频特征的映射关系,对参考帧的高频特征进行恢复和隐式的特征对齐与提取。然后,将两个模块得到的特征分别输入到注意力融合模块中,通过使用时间、通道和空间混合注意力,使模型能够更加充分地利用多维度信息,进而提升模型的特征表达能力。最后,通过对融合后的特征进行重建,得到高分辨率的视频帧。

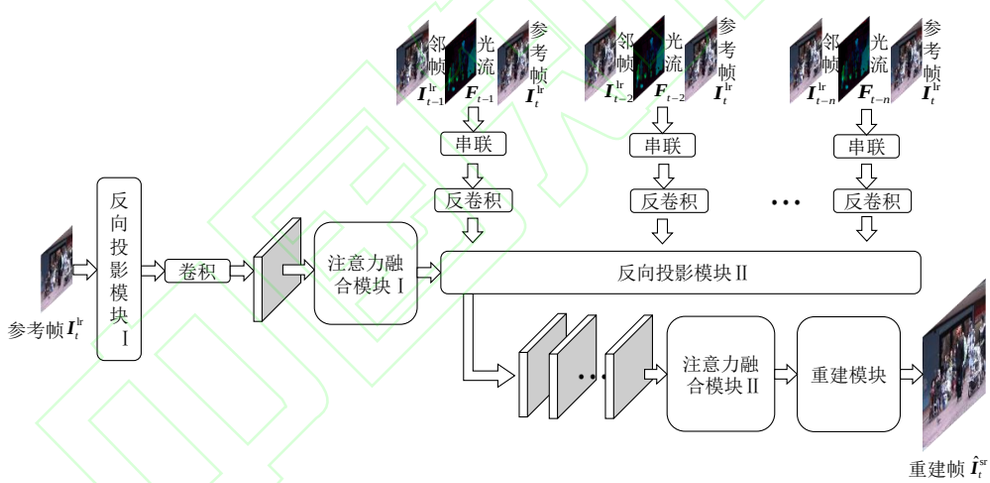


图2 注意力融合网络(AFN)整体框架

Fig. 2 Overall framework of Attention Fusion Network (AFN)

2.2 反向投影模块

反向投影机制是一种能够有效最小化重建损失的迭代过程[34],将反向投影机制应用于超分辨率主要是基于高分辨率图像下采样后得到的低分辨率图像应该与原来的低分辨率图像尽可能相似的假设,通过采用迭代的上采样和下采样方式去计算不同网络深度每个阶段的重建错误[35]。本文采用了两个反向投影模块,两个模块的整体处理流程一致,只是模块的输入不同。

第一个反向投影模块中,输入是经过双三次插值下采样得到的t时刻低分辨率参考帧$\boldsymbol{I}_t^{\mathrm{lr}}$,首先通过两个卷积层进行特征提取,得到初始的特征图$\boldsymbol{L}_t^0$:

$$\boldsymbol{L}_t^0=\mathrm{Conv}(\boldsymbol{I}_t^{\mathrm{lr}}) \tag{2}$$

其中Conv(·)表示卷积操作。

第i阶段的输入是前一个阶段$i-1,i\in\{1,...,n\}$的低分辨率特征图$\boldsymbol{L}_t^{i-1}$,采用上投影和下投影的结构,先将特征图依次通过反卷积操作和卷积操作进行上采样和下采样,得到的特征图与原来特征图大小相同,然后将两个特征图相减得到残差,即重建损失。第i阶段上投影过程得到的特征图为$\boldsymbol{H}_t^i$:

$$\boldsymbol{e}_i^{\mathrm{lr}}=\mathrm{Conv}(\mathrm{DConv}(\boldsymbol{L}_t^{i-1}))-\boldsymbol{L}_t^{i-1} \tag{3}$$

$$\boldsymbol{H}_t^i=\mathrm{DConv}(\boldsymbol{e}_i^{\mathrm{lr}})+\mathrm{DConv}(\boldsymbol{L}_t^{i-1}) \tag{4}$$

其中,$\boldsymbol{e}_i^{\mathrm{lr}}$表示第$i$个阶段上投影过程产生的残差,DConv(·)表示反卷积操作。

与上投影过程对应,下投影过程是将上投影得到的特征图$\boldsymbol{H}_t^i$下采样为特征图$\boldsymbol{L}_t^i$,作为下一阶段的输入特征:

$$\boldsymbol{e}_i^{\mathrm{hr}}=\mathrm{DConv}(\mathrm{Conv}(\boldsymbol{H}_t^i))-\boldsymbol{H}_t^i \tag{5}$$

$$\boldsymbol{L}_t^i=\mathrm{Conv}(\boldsymbol{e}_i^{\mathrm{hr}})+\mathrm{Conv}(\boldsymbol{H}_t^i) \tag{6}$$

其中,$\boldsymbol{e}_i^{\mathrm{hr}}$表示第$i$个阶段下投影过程产生的残差。

通过重复多个阶段的反向投影过程,不断迭代计算重建错误,最后将每个阶段通过反卷积放大得到的特征图串联在一起,经过卷积后得到特征$\boldsymbol{H}_t$:

$$\boldsymbol{H}_t = \mathrm{Conv}(\mathrm{Cat}(\boldsymbol{H}_t^1, \boldsymbol{H}_t^2, ..., \boldsymbol{H}_t^n)) \tag{7}$$

其中Cat(·)表示串联操作。

第二个反向投影模块中每个块的输入有两个，第一个输入是经过第一个注意力融合模块产生的特征图$\boldsymbol{H}_t^{\mathrm{lr}}$，第二个输入是通过将参考帧$\boldsymbol{I}_t^{\mathrm{lr}}$与单个邻帧$\boldsymbol{I}_{t-i}^{\mathrm{lr}}, i \in \{1, ..., n\}$以及相应的运动产生的光流$\boldsymbol{F}_{t-i}$串联在一起，经过反卷积操作后得到的特征图$\boldsymbol{H}_{t-i}^{\mathrm{m}}$：

$$\boldsymbol{H}_{t-i}^{\mathrm{m}} = \mathrm{DConv}(\mathrm{Cat}(\boldsymbol{I}_{t-i}^{\mathrm{lr}}, \boldsymbol{F}_{t-i}, \boldsymbol{I}_t^{\mathrm{lr}})) \tag{8}$$

然后采用和第一个投影模块类似的机制，不断迭代优化个阶段特征的差异$\boldsymbol{e}_{t-i}$，每个阶段产生的特征为$\boldsymbol{H}_{t-i}, i \in \{1, ..., n\}$，具体的实现过程如图3所示。

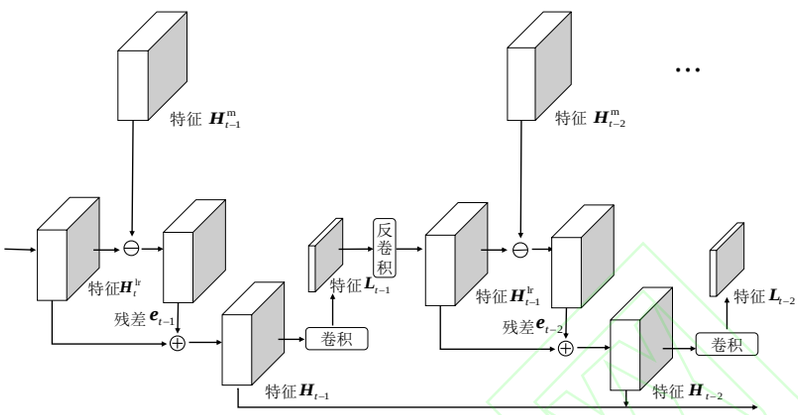


图3 反向投影模块Ⅱ中的上采样和下采样投影操作

Fig. 3 Up-sampling and down-sampling projection operation in back projection module Ⅱ

2.3 注意力融合模块

在2.2小节中提到的两个反向投影模块迭代产生了多个阶段的特征图，第一个反投影模块通过反向投影机制学习HR特征图和LR特征图之间的映射关系，并获取多阶段的误差反馈。第二个反投影模块由于结合了多个相邻帧的特征以及相邻帧与参考帧之间的运动而产生的光流信息，在保留低频信息的同时，能够获取一些运动信息并进行迭代的特征对齐，避免了显式的光流对齐操作，在减少运算量的同时可以较好地产生一些高频信息。但是如果只是将多个阶段的特征图简单融合在一起，并不能对这些特征充分利用，因为前后不同邻帧包含的信息量对参考帧的参考价值是不同的。例如，距离参考帧越远的邻帧与参考帧的特征相似度越低，而在视频的帧中存在运动模糊、遮挡、光线变化等问题时，需要同时考虑不同距离的邻帧，更加关注一些有益于参考帧重建的邻帧，而忽略掉一些存在问题的邻帧特征。同时，帧内特征的空间信息和通道信息也存在没有被充分的挖掘和利用的问题。

针对上面提到的视频超分辨率重建过程中存在的一些问题，本文提出了注意力融合模块。第一个模块在特征融合时采用了通道注意力和空间注意力，第二个模块考虑了时间信息，将时间注意力、通道注意力和空间注意力进行结合，共同指导特征的融合。与以往采用的空间注意力和通道注意力不同，本文采用了改进的空间注意力和通道注意力，分别称为增强空间注意力(ESA)和高效通道注意力(ECA)，通过融合多种注意力机制产生的特征，能够充分利用视频帧内和帧间信息，为邻帧特征和帧内特征分配不同权重，共同指导特征有选择地进行融合。在第二个注意力融合模块中，时间注意力、通道注意力和空间注意力的组合方式如图4所示，其中ECA表示高效通道注意力，ESA表示增强空间注意力。先通过时间注意力机制对多帧特征进行融合，然后使用通道注意力机制和空间注意力机制分别在通道维度和空间维度处理融合后的特征。

时间注意力机制用于计算参考帧的特征与多个邻帧特征之间的相似度。相似度越高，邻帧包含的信息量越多，对参考帧的重建参考价值越大。特征在嵌入空间的相似度计算可以使用卷积实现。首先将反向投影模块产生的特征$\{\boldsymbol{H}_{t-n}: \boldsymbol{H}_{t+n}\}$进行卷积，然后再将$\boldsymbol{H}_t$卷积后的特征分别与$\{\boldsymbol{H}_{t-n}: \boldsymbol{H}_{t-1}, \boldsymbol{H}_{t+1}: \boldsymbol{H}_{t+n}\}$卷积后的特征进行点积运算，最后通过Sigmoid激活函数使输出$\{\boldsymbol{A}_{t-n}: \boldsymbol{A}_{t+n}\}$的值在0到1之间，以使训练稳定，$\boldsymbol{A}_{t+i}$可以表示为：

$$\boldsymbol{A}_{t+i} = \sigma(\mathrm{Conv}(\boldsymbol{H}_t)\mathrm{Conv}(\boldsymbol{H}_{t+i})) \tag{9}$$

其中，$i \in [-n: +n]$，$\sigma(\cdot)$表示Sigmoid激活函数。

得到$\{\boldsymbol{A}_{t-n}: \boldsymbol{A}_{t+n}\}$之后，需要将其与原来的特征$\{\boldsymbol{H}_{t-n}: \boldsymbol{H}_{t-1}, \boldsymbol{H}_{t+1}: \boldsymbol{H}_{t+n}\}$按元素相乘，然后将经过时间注意力处理后的结果串联并进行一次额外的卷积操作得到融合后的特征$\boldsymbol{H}^{\mathrm{f}}$：

$$\boldsymbol{H}^{\mathrm{f}} = \mathrm{Conv}(\mathrm{Cat}(\boldsymbol{H}_{t-n} \odot \boldsymbol{A}_{t-n}, ..., \boldsymbol{H}_{t+n} \odot \boldsymbol{A}_{t+n})) \tag{10}$$

其中，$\odot$表示按元素相乘。

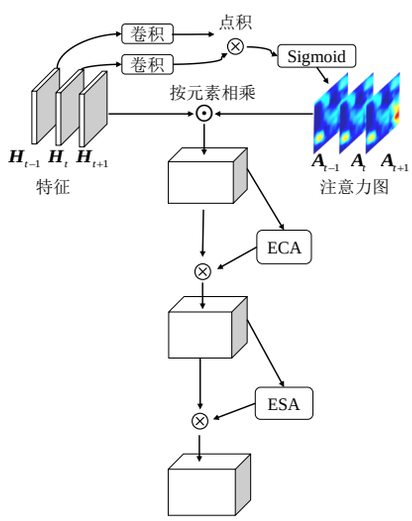


图4 注意力融合模块Ⅱ

Fig. 4 Attention fusion module Ⅱ

得到融合特征$\boldsymbol{H}^{\mathrm{f}}$后,将其作为ECA模块的输入。在ECA模块中,先将$\boldsymbol{H}^{\mathrm{f}}$进行全局池化(Global Pooling, GP)来获取全局上下文,包括全局平均池化(Global Average Pooling, GAP)和全局最大池化(Global Max Pooling, GMP),然后由通道维度的映射自适应选择一维卷积核的大小k并经过卷积生成通道权重,这样可以在没有维度约简的情况下实现跨通道交互,在本文中,k和$\boldsymbol{H}^{\mathrm{f}}$的通道数$C$的关系为:

$$k=\left|\frac{\log_2 C}{2}+\frac{1}{2}\right|_{\mathrm{odd}} \tag{11}$$

其中,$|\cdot|_{\mathrm{odd}}$表示取最接近的奇数。

通过将ECA融入本文的注意力模块中,能够高效利用特征的通道信息,提升深度网络对视频帧进行重建的性能。图5是通道注意力和ECA之间的结构对比。

在增强空间注意力模块中,如图6所示,为了在获得更大的感受野的同时能够轻量化,首先将经过通道注意力处理后的特征通过1×1卷积降低通道维度,然后采用跨步卷积和全局平均池化来获得更大的感受野。在经过一组卷积以后,采用上采样和1×1卷积恢复空间和通道维度,同时,使用残差连接将通道缩减之前的特征与上采样之后的特征相加,最后使用Sigmoid激活函数得到空间权重。

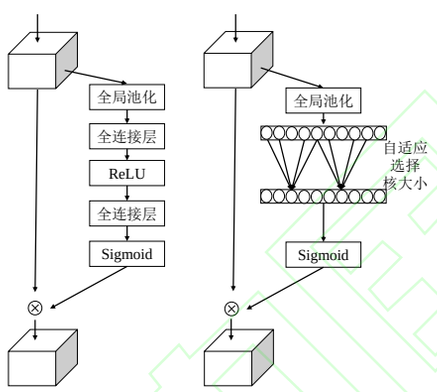


图5 通道注意力和高效通道注意力(ECA)的对比

Fig. 5 Comparison of channel attention and efficient channel attention (ECA)

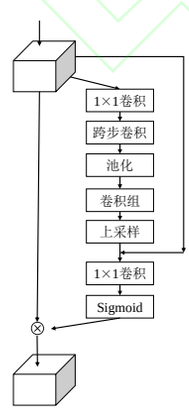


图6 增强空间注意力(ESA)模块

Fig. 6 Enhanced spatial attention (ESA) module

与非局部注意力相比,增强空间注意力在获取长范围特征依赖的同时更加轻量化,因此也可以将多个ESA模块分别放在多个阶段反向投影模块的残差块后面,从而实现让残差特征更加关注重要的空间信息以及能够获得更大的感受野。

在卷积块注意力模块(Convolutional Block Attention Module, CBAM)[36]网络中,讨论了通道注意力模块和空间注意力模块连接的顺序和方式对模型性能的影响。通过相关实验证明,两个模块按顺序连接产生的注意力图比并联方式产生的注意力图更加精细,并且先使用通道注意力模块比先使用空间注意力模块的性能表现稍好。借鉴CBAM中注意力模块的连接方式和连接顺序,本文中第一个反向投影模块产生的特征以及时间注意力模块融合后的特征,先进入高效通道注意力模块进行处理,再进入增强空间注意力模块,两个注意力模块按顺序连接。

3 相关实验

3.1 数据集与实现细节

本文采用Vimeo-90K[37]作为训练数据集,Vimeo-90K是一个大规模、高质量的视频数据集,包含了多种场景和运动,该数据集总共包括64612个7帧视频序列,每帧的大小为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,并对视频帧使用了随机裁剪、旋转等数据扩增技术。测试数据集采用Vid4[9]、SPMCS[18]。Vid4数据集共有4个不同场景的视频片段,包含多种运动和遮挡,SPMCS数据集总共包含30个不同的视频场景,每个视频有31帧。对于4倍超分辨率,使用的输入帧是112×64大小的低分辨率视频帧进行裁剪得到的64×64大小的图像块,批次大小为8。

在第一个反向投影模块,首先通过两个卷积层进行初始的参考帧特征提取,经过第一个卷积层后特征通道数为256,经过第二个卷积层通道数变为64,然后使用了编码解码的结构,包含卷积和反卷积操作用于对特征图进行上采样和下采样,总共迭代了三次。第二个反向投影模块使用了残差模块进行特征提取以及使用反卷积进行上采样,每个残差模块包含5个残差块。所有的卷积和反卷积操作的参数都使用了Kaiming初始化[38]方法。在第二个注意力融合模块中,将反投影模块中各个阶段的特征堆叠,首先使用时间注意力对多个高分辨率特征进行融合,然后使用高效通道注意力,最后使用增强空间注意力,经过注意力融合模块后的高分辨率特征通道数为64,需要再经过一个卷积层恢复3个通道,最后得到高分辨率的目标帧。由于经过反卷积层生成的特征图大小与目标帧的大小一致,因此不需要再进行上采样操作。本文在Ubuntu 18.04操作系统上使用PyTorch框架实现了AFN,在4块Nvidia Tesla V100显卡上进行模型的训练,训练时使用了L1损失函数,定义为:

$$L_{\mathrm{sr}}=\frac{1}{N}\sum_{t=1}^{N}\left\|\boldsymbol{I}_t^{\mathrm{hr}}-\hat{\boldsymbol{I}}_t^{\mathrm{sr}}\right\|_1 \tag{12}$$

其中$\boldsymbol{I}_t^{\mathrm{hr}}$和$\hat{\boldsymbol{I}}_t^{\mathrm{sr}}$分别表示$t$时刻的高分辨率目标帧和重建的视频帧,$N$为训练数据总帧数。优化器为Adam[39],动量为0.9,学习率初始化为1e-4。总共训练了150个轮次,在总轮次的一半时学习率下降为原来的十分之一。每10个轮次保存一次模型参数。

3.2 与主流方法的对比实验

实验中对视频帧进行4倍超分辨率重建,使用峰值信噪比(Peak Signal-to-Noise Ratio, PSNR)[40]和结构相似性(Structural SIMilarity, SSIM)[41]作为量化指标来度量重建视频

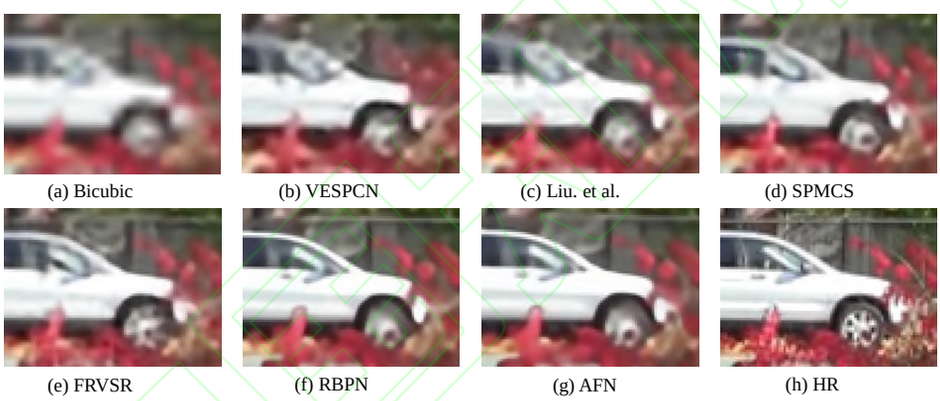
帧的质量，PSNR 基于对应像素点间的误差衡量图像质量，SSIM 分别从亮度、对比度和结构三方面度量图像相似性，PSNR 和 SSIM 的指标越高，表示重建视频帧的质量越好。分别在 Vid4 和 SPMCS 数据集上对视频序列在 YUV 颜色空间中的 Y（亮度）通道上进行测试，并与一些主流的视频超分辨率重建方法进行了对比实验。同时，为了验证提出的注意力融合模块的有效性，本文进行了相关的消融实验。

首先通过对 Vid4 数据集上的四个视频片段进行测试，比较了 AFN、VESPCN、SPMC、FRVSR 和 RBPN 方法的性能。根据以往的指标计算方法，在测试时去除了每个视频序列中的前两帧和后两帧。测试的结果如表 1 所示。通过分析在 Vid4 数据集上每个视频序列测试的结果以及最后的平均指标，可以看到，与其他视频超分辨率重建方法相比，本文的方法整体上在 PSNR 和 SSIM 指标方面有所提高。图 7 展示了不同方法对 Vid4 数据集中的 foliage 视频片段进行 4 倍超分辨率后的视觉效果对比，可以看到 VESPCN 和 SPMCS 等方法产生的图像过于模糊且存在一定程度的结构失真，本文的方法产生的视频帧包含更少的噪声，纹理方面恢复得较好。虽然 FRVSR 方法在恢复车轮的结构细节时处理得比较好，但是图像整体上比较模糊，且包含一些噪声，使得图像不够清晰。

表 1　在 Vid4 数据集上进行 4 倍超分辨率的 PSNR 和 SSIM 结果

Tab. 1　PSNR and SSIM results on Vid4 dataset for 4×super-resolution

视频片段	Bicubic		VESPCN		SPMC		FRVSR		RBPN		AFN（Ours）	
	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM
Calendar	19. 82	0. 554	-	-	22. 16	0. 746	-	-	**23. 99**	**0. 807**	23. 92	0. 807
City	24. 93	0. 586	-	-	27. 00	0. 757	-	-	27. 73	0. 803	**27. 75**	**0. 804**
Foliage	23. 42	0. 575	-	-	25. 43	0. 721	-	-	26. 22	0. 757	**26. 28**	**0. 759**
Walk	26. 03	0. 802	-	-	28. 91	0. 876	-	-	30. 70	0. 909	**30. 74**	**0. 920**
Average	23. 53	0. 629	25. 35	0. 757	25. 88	0. 775	26. 69	0. 822	27. 12	0. 818	**27. 17**	**0. 822**



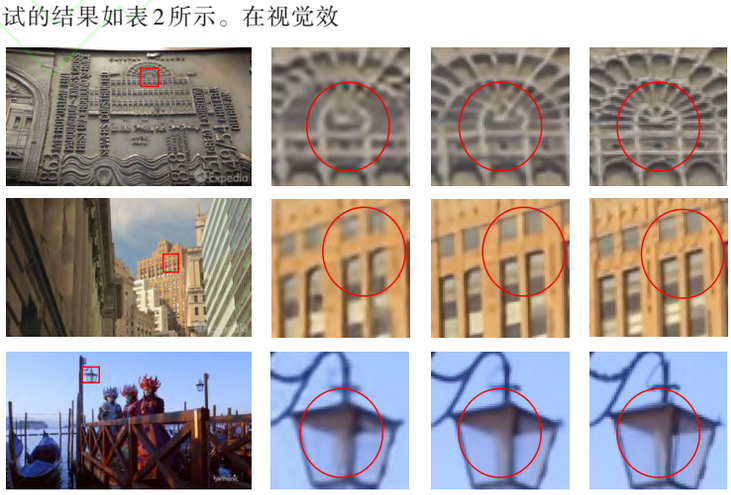
(a) Bicubic　(b) VESPCN　(c) Liu. et al.　(d) SPMCS

(e) FRVSR　(f) RBPN　(g) AFN　(h) HR

图 7　在 Vid4 数据集上对 foliage 视频片段进行 4 倍超分辨率的视觉效果对比

Fig. 7　Visual comparison on foliage video clip of Vid4 dataset for 4×super-resolution

在 SPMCS 数据集上，选取了 11 个视频片段用于比较 AFN、SPMC、VSR-DUF 和 RBPN 方法进行 4 倍超分辨率重建的性能表现。在计算 PSNR 和 SSIM 指标时，去除了每个视频序列中的前 6 帧和后 3 帧。测试的结果如表 2 所示。在视觉效果方面，与 SPMC 相比，如图 8 所示，本文的方法产生的视频帧在物体的边缘纹理方面更加清晰，能够更好地恢复视频帧内的细节和结构信息。



(a) SPMC　(b) AFN　(c) HR

图 8　在 SPMCS 数据集上进行 4 倍超分辨率的视觉效果对比

Fig. 8　Visual comparison on SPMCS dataset for 4×super-resolution

通过以上的对比实验可以发现,与一些主流的视频超分辨率重建方法相比,本文方法重建的视频帧在客观评价指标和主观视觉效果方面有进一步的提升。

表2 在SPMCS数据集上进行4倍超分辨率的PSNR结果单位:dB

Tab. 2 PSNR results on SPMCS dataset for 4×super-resolutionunit: dB

视频片段	Bicubic	SPMC	VSR-DUF	RBPN	AFN
car05_001	27.62	**32.07**	30.77	31.65	31.90
hdclub_003_001	19.38	21.03	**22.07**	21.91	21.90
hitachi_isee5_001	19.59	23.83	25.73	26.14	**26.32**
hk004_006	28.46	32.14	32.96	**33.25**	33.07
HKVTG_004	27.37	28.71	29.15	29.39	**29.46**
jvc_009_001	25.31	28.15	29.26	30.17	**30.23**
NYVTG_006	28.46	31.46	32.29	33.09	**33.18**
PRVTG_012	25.54	26.95	27.47	27.52	**27.63**
RMVTG_011	24.00	26.49	27.63	27.64	**27.70**
veni3_011	29.32	34.66	34.51	36.14	**36.60**
veni5_015	27.30	31.75	32.45	32.66	**33.12**
Average	25.67	28.82	29.42	29.96	**30.10**

3.3 消融实验

为了验证本文提出的注意力融合机制的有效性,将移除了注意力融合模块的网络作为基线模型并进行模型训练与测试。相关的实验结果如表3所示,其中w/和w/o分别表示有和没有注意力融合模块。

与基线模型相比,AFN在Vid4和SPMCS数据集上重建视频帧的PSNR值分别平均提高了0.24 dB和0.31 dB,SSIM值分别平均提高了0.011和0.024,证明了本文提出的注意力融合机制在超分辨率重建过程中的有效性。

表3 注意力融合对AFN重建视频帧PSNR和SSIM的影响

Tab. 3 Effect of attention fusion on PSNR and SSIM of video frames reconstructed by AFN

数据集	w/		w/o	
	PSNR/dB	SSIM	PSNR/dB	SSIM
Vid4	**27.17**	**0.822**	26.93	0.811
SPMCS	**30.10**	**0.881**	29.79	0.857

消融实验的结果也说明在对特征进行融合时,帧内的空间信息、通道信息和帧间的时间信息应该可以被更好地进行探索和利用。

4 结语

本文提出了一个视频超分辨率重建模型,称为注意力融合网络(AFN)。通过采用反向投影原理,结合多种最新的注意力机制和融合策略,AFN能够更好地利用视频帧内和帧间信息,充分挖掘帧内特征和帧间特征的相关关系,从而有效处理包含多种运动和遮挡的视频。通过进行相关的对比实验,与一些主流的视频超分辨率重建方法相比,本文模型具有更优的客观性能指标和更好的视觉效果。通过消融实验,验证了部分网络模块的有效性。在未来的工作中,我们将致力于不断改进网络模型结构,在提高网络模型性能的同时,使模型更加轻量化。

参考文献(References)(References)

[1] DONG C, LOY C C, HE K, et al. Image super-resolution using deep convolutional networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 28(2)295-307.

[2] KIM J, LEE J K. LEE K M. Accurate image super-resolution using very deep convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE, 2016: 1646-1654.

[3] LIM B, SON S, KIM H, et al. Enhanced deep residual networks for single image super-resolution[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE, 2017: 1132-1140.

[4] ZHANG Y, TIAN Y, KONG Y, et al. Residual dense network for image super-resolution//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE, 2018: 2472-2481.

[5] DAI T, CAI J, ZHANG Y, et al. Second-order attention network for single image super-resolution//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE, 2019: 11057-11066.

[6] GUO Y, CHEN J, WANG J, et al. Closed-loop matters: dual regression networks for single image super-resolution//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE, 2020: 5406-5415.

[7] GARCIA D C, DOREA C, DE QUEIROZ R L. Super resolution for multiview images using depth information[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 22(9): 1249.

[8] FARAMARZI E, RAJAN D, CHRISTENSEN M P. Unified blind method for multi-image super-resolution and single/multi-image blur deconvolution[J]. IEEE Transactions on Image Processing, 2013, 22(6): 2101-2114.

[9] LIU C, SUN D. A bayesian approach to adaptive video super resolution[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2011: 209-216.

[10] GUO J, CHAO H. Building an end-to-end spatial-temporal convolutional network for video super-resolution[C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2017: 4053-4060.

[11] SAJJADI M S M, VEMULAPALLI R, BROWN M. Frame-recurrent video super-resolution[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 6626-6634.

[12] WANG X, CHAN K C K, YU K, et al. Edvr: video restoration with enhanced deformable convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2019: 1954-1963.

[13] 何小海, 吴媛媛, 陈为龙, 等. 视频超分辨率重建技术综述[J]. 太赫兹科学与电子信息学报, 2011, 9(1): 1-6. (HE X, WU Y Y, CHEN W L, et al. Review of video super resolution reconstruction technology[J]. Journal of Terahertz Science and Electronic Information Technology, 2011, 9(1): 1-6.)

[14] ANWAR S, KHAN S, BARNES N. A deep journey into super-resolution: a survey[J]. ACM Computing Surveys, 2020, 53(3): 1-34.

[15] DONG C, LOY C C, HE K, et al. Learning a deep convolutional network for image super-resolution[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2014: 184-199.

[16] KAPPELER A, YOO S, DAI Q, et al. Video super-resolution with convolutional neural networks[J]. IEEE Transactions on

Computational Imaging, 2016, 2(2):109-122.
[17] CABALLERO J, LEDIG C, AITKEN A, et al. Real-time video super-resolution with spatio-temporal networks and motion compensation [J]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 2848-2857.
[18] TAO X, GAO H, LIAO R, et al. Detail-revealing deep video super-resolution [J]//Proceedings of the IEEE Conference on Computer Vision. Piscataway: IEEE, 2017: 22-29.
[19] SHI X, CHEN Z, WANG H, et al. Convolutional lstm network: a machine learning approach for precipitation nowcasting [C]// Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2015: 802-810.
[20] LIU D, WANG Z, FAN Y, et al. Robust video super-resolution with learned temporal dynamics [C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2526-2534.
[21] JO Y, OH S W, KANG J, et al. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 3224-3232.
[22] D. TRAN, L. BOURDEV, R. FERGUS, et al. Learning spatiotemporal features with 3d convolutional networks [C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2015: 4489-4497.
[23] HARIS M, SHAKHNAROVICH G, UKITA N. Recurrent back-projection network for video super-resolution [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 3892-3901.
[24] YI P, WANY Z, JIANG K, et al. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 3106-3115.
[25] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7794-7803.
[26] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer networks [C]//Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2015: 2017-2025.
[27] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7132-7141.
[28] WANG Q, WU B, ZHU P, et al. Eca-net: efficient channel attention for deep convolutional neural networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 11531-11539.
[29] WANG F, JIANG M, QIAN C, et al. Residual attention network for image classification [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 6450-6458.
[30] LIU Y, WANG Y, LI N, et al. An attention-based approach for single image super resolution [C]//24th International Conference on Pattern Recognition [C]. Piscataway: IEEE, 2018: 2777-2784
[31] ZHANG Y, LI K, LI K, et al. Image super-resolution using very deep residual channel attention networks [C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 286-301.
[32] LIU Z, WANG L, LI C, et al. Image super-resolution via attention based back projection networks [C]//Proceedings of the IEEE International Conference on Computer Vision Workshop. Piscataway: IEEE, 2019: 3517-3525.
[33] LIU J, ZHANG W, TANG Y, et al. Residual feature aggregation network for image super-resolution [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 2356-2365.
[34] IRANI M, PELEG S. Improving resolution by image registration [J]. CVGIP: Graphical models and image processing, 1991, 53 (3):231 - 239.
[35] HARIS M, SHAKHNAROVICH G, UKITA N. Deep back-projection networks for single image super-resolution [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 1664-1673.
[36] WOO S, PARK J, LEE JY, et al. Cbam: convolutional block attention module [C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 3-19.
[37] XUE T, CHEN B, WU J, et al. Video enhancement with task-oriented flow [J]. International journal of computer vision, 2017: 10. 1007.
[38] HE K, ZHANG X, REN S, et al. Delving deep into rectifiers: surpassing human-level performance on imagenet classification [C]//Proceedings of the IEEE Conference on Computer Vision. Piscataway: IEEE, 2015: 1026-1034.
[39] DIEDERIK K, JIMMY B. Adam: a method for stochastic optimization [EB/OL]. [2018-12-22]. https://arxiv. org/pdf/1412. 6980. pdf. .
[40] 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.
[41] ALAIN H, DJEMEL Z. Image quality metrics: Psnr vs. Ssim [C]// International Conference on Pattern Recognition. Piscataway: IEEE, 2010: 2366-2369.

This work is partially supported by the National Natural Science Foundation of China (61672467).

BIAN Pengcheng, born in 1993, M. S. candidate. His research interests include deep learning, computer vision.;

ZHENG Zhonglong, born in 1976, Ph. D., professor. His research interests include pattern recognition, machine learning, image processing.;

LI Minglu, born in 1965, Ph. D., professor. His research interests include cloud computing, vehicular ad hoc network, wireless sensor network, big data analysis.;

HE Yiran, born in 1996, M. S.. Her research interest includes machine learning.;

WANG Tianxiang, born in 1994, Ph. D. candidate. His research interests include machine learning, computer vision.;

ZHANG Dawei, born in 1995, Ph. D. candidate. His research interests include deep learning, computer vision.;

CHEN Liyuan, born in 1994, Ph. D. candidate. Her research interests include deep learning, computer vision.