中图法分类号:TP391　　文献标识码:A　　文章编号:1006-8961(2023)07-1927-38

论文引用格式:Jiang J J, Cheng H, Li Z Y, Liu X M and Wang Z Y. 2023. Deep learning based video-related super-resolution technique: a survey. Journal of Image and Graphics, 28(07):1927-1964(江俊君, 程豪, 李震宇, 刘贤明, 王中元. 2023. 深度学习视频超分辨率技术综述. 中国图象图形学报, 28(07):1927-1964)[DOI:10.11834/jig.220130]

深度学习视频超分辨率技术综述

江俊君[1*],程豪[1],李震宇[1],刘贤明[1],王中元[2]

1. 哈尔滨工业大学计算机科学与技术学院,哈尔滨　150001; 2. 武汉大学计算机学院,武汉　430072

摘　要: 视频超分辨率技术在卫星遥感侦测、视频监控和医疗影像等方面发挥着关键作用,在各领域具有广阔的应用前景,受到广泛关注,但传统的视频超分辨率算法具有一定局限性。随着深度学习技术的愈发成熟,基于深度神经网络的超分辨率算法在性能上取得了长足进步。充分融合视频时空信息可以快速高效地恢复真实且自然的纹理,视频超分辨率算法因其独特的优势成为一个研究热点。本文系统地对基于深度学习的视频超分辨率的研究进展进行详细综述,对基于深度学习的视频超分辨率技术的数据集和评价指标进行全面归纳,将现有视频超分辨率方法按研究思路分成两大类,即基于图像配准的视频超分辨率方法和非图像配准的视频超分辨率方法,并进一步立足于深度卷积神经网络的模型结构、模型优化历程和运动估计补偿的方法将视频超分辨率网络细分为10个子类,同时利用充足的实验数据对每种方法的核心思想以及网络结构的优缺点进行了对比分析。尽管视频超分辨率网络的重建效果在不断优化,模型参数量在逐渐降低,训练和推理速度在不断加快,然而已有的网络模型在性能上仍然存在提升的潜能。本文对基于深度学习的视频超分辨率技术存在的挑战和未来的发展前景进行了讨论。

关键词: 深度学习;视频超分辨率(VSR);图像配准;运动估计;运动补偿

Deep learning based video-related super-resolution technique: a survey

Jiang Junjun[1*], Cheng Hao[1], Li Zhenyu[1], Liu Xianming[1], Wang Zhongyuan[2]

1. School of Computer Science and Technology, Harbin Institute of Technology, Harbin　150001, China;

2. School of Computer, Wuhan University, Wuhan　430072, China

Abstract: Video-related super-resolution (VSR) technique can be focused on high-resolution video profiling and restoration to optimize its low-resolution version-derived quality. It has been developing intensively in relevant to such domains like satellite remote sensing detection, video surveillance, medical imaging, and low-involved electronics. To reconstruct high-resolution frames, conventional video-relevant super-resolution methods can be used to estimate potential motion status and blur kernel parameters, which are challenged for multiscene hetegerneity. Due to the quick response ability of fully integrating video spatio-temporal information of real and natural textures, the emerging deep learning based video super-resolution algorithms have been developing dramatically. We review and analyze current situation of deep learning based video super-resolution systematically and literately. First, popular YCbCr datasets are introduced like YUV25, YUV21, ultra video group(UVG), and the RGB datasets are involved in as well, such as video 4 (Vid4), realistic and dynamic scenes (REDS), Vimeo90K. The profile information of each dataset is summarized, including its name, year of publication, number of videos, frame number, and resolution. Furthermore, key parameters of the video super-resolution algo-

收稿日期:2022-02-28;**修回日期:**2022-05-21;**预印本日期:**2022-05-28

*通信作者:江俊君　jiangjunjun@hit.edu

基金项目:国家自然科学基金项目(61971165,92270116,62071339)

Supported by: National Natural Science Foundation of China (61971165, 92270116, 62071339)

rithm are introduced in detail in terms of peak signal-to-noise ratio (PSNR), structural similarity (SSIM), video quality model for variable frame delay (VQM_VFD), and learned perceptual image patch similarity (LPIPS). For the concept of video super-resolution and single image super-resolution, the difference between video super-resolution and single image super-resolution can be shown and the former one has richer video frames-interrelated motion information. If the video is processed frame by frame in terms of the single image super-resolution method, there would be a large number of artifacts in the reconstructed video. We carry out deep learning based video super-resolution methods analysis and it has two key technical challenges of those are image alignment and feature integration. For image alignment, its option of image alignment module is challenged for severe hetergeneity between video super-resolution methods. Image alignment and non-alignment methods are categorized. The integration of multi-frame information is based on the network structure like generative adversarial networks (GAN), recurrent convolutional neural networks (RNN), and Transformer. To process video feature and make neighboring frames align with the target frame, image-aligned methods can use different motion estimation and motion compensation module. Image alignment methods can be segmented into three alignment-related categories: optical flow, kernel, and convolution-deformable. This optical flow alignment method can be used to calculate the motion flows between two frames through their pixels-between gray changes in temporal and the neighboring frames are warped by motion compensation module. We divide them into four categories in terms of the optical flow alignment-relevant model structure of deep convolutional neural network (CNN) further: 2D convolution, RNN, GAN, and Transformer. For optical flow-aligned 2D convolution methods analysis, we mainly introduce video efficient sub-pixel convolutional network (VESPCN) and its improvement on optical flow estimation network and motion compensation network, such as ToFlow and spatial-temporal transformer network (STTN). For the RNN methods with optical flow alignment, we analyze residual recurrent convolutional network (RRCN), recurrent back-projection network (RBPN) and other related methods using optical flow to align neighboring frames at the image level, which is required to resolve the constraints of the sliding window methods. Therefore, to obtain excellent reconstruction performance, we focus on BasicVSR (basic video super-resolution), IconVSR (information-refill mechanism and coupled propagation video super-resolution) and other networks, which can warp neighboring frames at the feature level. The optical flow alignment-based TecoGAN (temporal coherence via self-supervision for gan-based video generation) and VSR Transformer methods are introduced in detail as well. Due to a few kernel-based and deformable convolution-based align methods, it is still a challenging issue for classify network structure. Because convolution kernel size can used to limit the range of motion estimation, the reconstruction performance of the kernel-based alignment methods is relatively poor. Specifically, deformable convolution is a sampling improvement of conventional convolution, which still has some gaps to be bridged like high computational complexity and harsh convergence conditions. For non-alignment methods, multiple network structures are challenged for video frames-between correlation to a certain extent. We review and analyze the methods in related to non-aligned 3D convolution, non-aligned RNN, alignment-excluded GAN, and non-local. The non-alignment RNN methods consist of recurrent latent space propagation (RLSP), recurrent residual network (RRN) and omniscient video super-resolution (OVSR) and it demonstrates that a balance can be achieved between reconstruction speed and visual quality. To reduce the computational cost, the improved non-local module is focused on when alignment-excluded non-local methods are introduced. All models are tested with 4× downsampling using two degradations like bicubic interpolation (BI) and blur downsampling (BD). The multiple datasets-based quantitative results, speed comparison of the super-resolution methods are summarized as well, including REDS4, UDM10, and Vid4. Some effects can be optimized. The reconstruction performances of these video-based super-resolution networks are balanced in consistency, the parameters of the model are gradually shrinked, and the speed of training and reasoning is accelerated as well. However, the application of deep learning in video super-resolution is still to be facilitated more. We predict that it is necessary to improve the adaptability of the network and validate the traced result. Current deep learning technologies can be introduced on the nine aspects as mentioned below: network training and optimization, ultra-high resolution-oriented video super-resolution for, video-compressed super-resolution video-rescaling methods, self-supervised video super-resolution, various-scaled video super-resolution, spatio-temporal video super-resolution, auxiliary task-guided video super-resolution, and scenario-customized video super-resolution.

Key words: deep learning; video super-resolution (VSR); image alignment; motion estimation; motion compensation

0 引　言

超分辨率技术(super-resolution,SR)是近年来计算机视觉和图像处理领域中的一个研究热点,其主要目标是将低分辨率图像/视频转换为高分辨率图像/视频。随着时代的发展,人们对图像/视频有了更高质量的要求。在传输过程中,由于传输带宽的限制,需要将图像/视频进行压缩处理,因此存在细节信息的损失,需要使用相应的手段进行质量提升。在安全防范方面,视频监控因环境、设备等因素的影响,获得的视频分辨率较低。超分辨率技术应用于监视分析(Zhang等,2010)和人脸识别(Gunturk等,2003;Li等,2018),可以获取车牌和人脸等感兴趣对象的关键信息。在医疗领域,超分辨率技术通过恢复医疗影像的特征细节来帮助病患获得及时准确的治疗。

超分辨率技术主要分为传统方法和基于深度学习的方法两类。基于插值的超分辨率方法实现简单,且已得到广泛应用。如最近邻插值法、双线性插值法和双三次插值法(Dong等,2016;Niklaus等,2017a)等,通过利用周围像素点的像素值估计丢失的像素值,但是这些线性模型限制了它们恢复高频能力的细节。1964年,Harris(1964)提出了Harris-Goodman频谱外推法。1984年,Huang和Tsai(1984)利用傅里叶变换来实现高分辨率图像的重建。Liu和Sun(2014)提出一种贝叶斯方法来估计帧间运动,进而重建高分辨率视频。这些算法受限于特定假设,在满足条件的情况下能够获得较好的仿真结果。但由于实际场景无法满足所有假设条件,因此此类算法在实际应用中的效果并不理想。

随着高分辨率显示设备的普及,传统的超分辨率方法由于难以恢复视频和图像中的高频细节信息,无法满足4倍放大因子的需求。而深度学习在图像、语音、视频以及自然语言等领域取得的巨大成功为超分辨率算法研究提供了新的思路。基于深度学习的超分辨率算法展现出了强大的非线性学习能力,逐渐取代了传统超分辨率技术。

目前,基于深度学习的图像超分辨率算法的综述较多(Wang等,2021;Singh和Singh,2020;Yang等,2019;李书林 等,2022),而基于深度学习的视频超分辨率的文献综述工作尚未引起广泛关注。吴洋和樊桂花(2017)、何小海等人(2011)对基于传统方法的视频超分辨率进行了综述,其中关于深度学习的方法涉及较少。Liu等人(2022)对基于深度学习的视频超分辨率方法进行了综述,但对各方法的实验总结分析较少,大多是对网络结构和训练方法的阐述。近两年,视频超分辨率迅速发展,RLSP(recurrent latent space propagation)(Fuoli等,2019)、RRN(recurrent residual network)(Isobe等,2020c)和OVSR(omniscient video wuper-resolution)(Yi等,2021)等方法通过设计轻量级网络可以实时重建高分辨率。BasicVSR(basic video wuper-resolution)(Chan等,2021a)、IconVSR(information-refill mechanism and coupled propagation video super-resolution)(Chan等,2021a)和BasicVSR++(Chan等,2021b)等在性能上实现了重大突破,展现出了惊人的结果。Transformer在CV(computer vision)领域大放异彩,在视频超分辨率领域迅速得到广泛关注,然而这些方法在Liu等人(2022)的综述中均未提及。深度学习的视频超分辨率研究仍在进一步发展,本文对该领域既有的研究成果进行概述提炼,希望为视频超分辨率重建相关领域的研究提供参考和帮助。

本文结构如图1所示。基于深度学习的视频超分辨率方法存在两大关键技术挑战,即相邻帧的配准和多帧信息的融合重建。图像配准模块的选用与否是区分现有方法的主要因素,因此本文从基于图像配准的方法和非图像配准方法两个方面,全面综述近年来视频超分辨率方法的现状。多帧信息的融合重建方式则取决于网络结构,视频超分辨率网络的具体分类如图2所示。对于基于图像配准的方法,本文将其分为3类,即基于光流的方法、基于卷积核估计的方法和基于可变形卷积的方法。并且,进一步将基于光流的方法细粒度划分为4种,即基于常规卷积的光流配准方法、基于生成对抗网络(generative adversarial network, GAN)的光流配准方法、基于循环神经网络的光流配准方法和基于Transformer的光流配准方法。对于非图像配准的视频超分辨率方法,本文将其分为4种,即基于3D卷积的非图像配准方法、基于GAN的非图像配准方法、基于循环神经网络的方法和基于non-local的非图像配准方法。

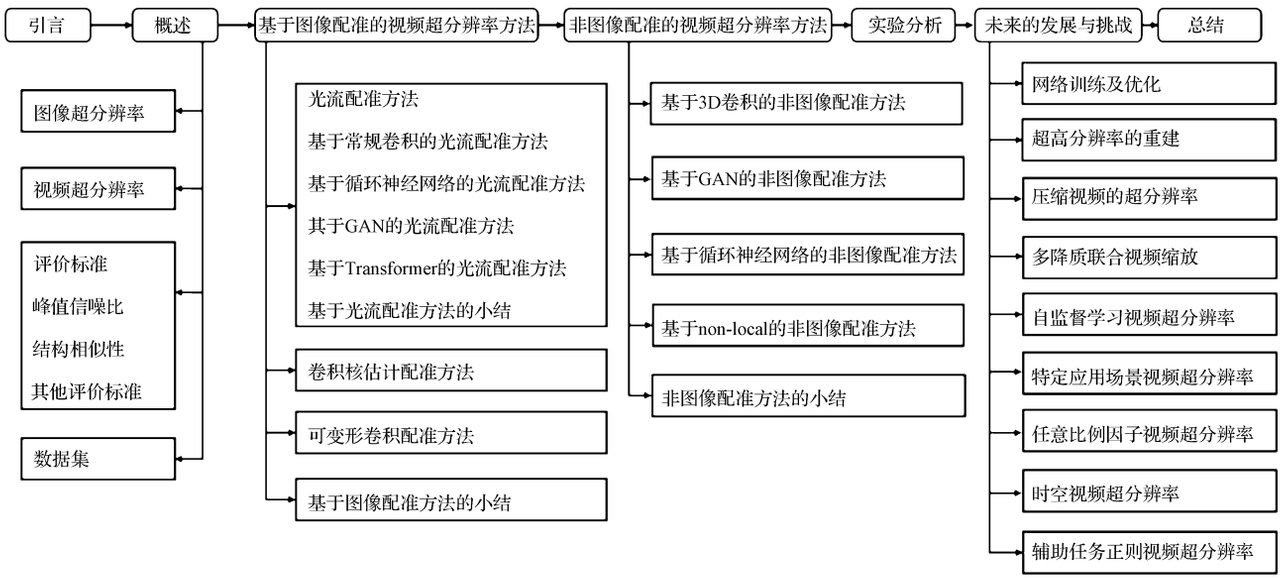

图1　本文架构

Fig. 1　The architecture of this paper

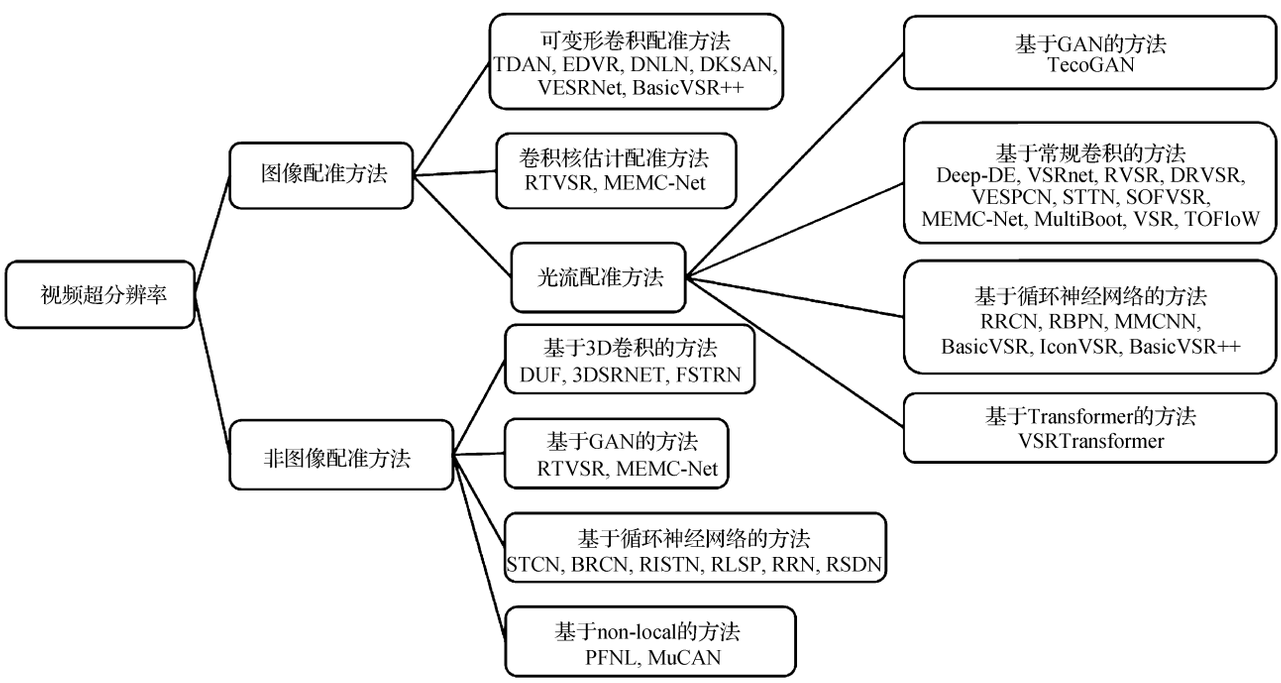

图2　基于深度学习的视频超分辨率算法的分类图

Fig. 2　A taxonomy for video super-resolution methods based on deep learning

1　概　述

1.1　图像超分辨率

图像超分辨率(Banham和Katsaggelos,1997)旨在通过1幅或多幅低分辨率图像获得高分辨率图像。卷积神经网络(convolutional neural networks,CNN)在许多计算机视觉任务中取得了良好效果。Dong等人(2014)首次将深度学习运用于图像超分辨率问题,将双三次插值法与卷积神经网络相结合提出SRCNN(super resolution CNN)。Shi等人(2016)设计的ESPCN(efficient sub-pixel CNN)取得了出色的性能,其中的亚像素卷积广泛应用于后续图像超分辨率网络的上采样模块。

残差网络(residual network)(He等,2016)缓解了网络训练时不易收敛和梯度消失的问题,可以很

好地扩充网络的深度，并使网络拥有更好的学习能力。Kim 等人（2016）将该方法运用于 VDSR（super-resolution using very deep convolutional network）网络进行图像超分辨率的研究，Zhang 等人（2018b）提出基于残差网络的 RCAN（residual channel attention network）。在扩展网络深度的同时，另一类研究是尝试增加网络内部各模块之间的连接，2017 年，Huang 等人（2017）提出密集连接网络（densely connected convolutional network，DenseNet）。DenseNet 不仅加强了网络内部的信息流动性，而且很大程度上减少了参数的使用。Tong 等人（2017）提出 SRDenseNet（super resolution DenseNet），首次将密集连接结构引入图像超分辨率研究。在以上研究基础上，Xu 等人（2018）将残差网络和密集网络相结合的残差密集块应用于图像超分辨率网络，最终生成细节真实的高分辨率图像。程德强等人（2021）利用递归方法将残差网络块进行复用，使用 32 层递归网络，增加网络深度并获取更丰富的特征信息。

随着 GAN（Goodfellow 等，2014）的发展，涌现出大量基于 GAN 的图像超分辨率方法。经典方法包括 SRGAN（super-resolution GAN）（Ledig 等，2017）、ProGAN（progressive growing of GANs）（Wang 等，2018b）、SFT-GAN（spatial feature transform GAN）（Wang 等，2018c）和 FSRNet（face super-resolution network）（Chen 等，2018）等。当放大因子为 4～8 时，这些方法的视觉重建效果好于大部分非对抗训练的图像超分辨率技术。近年来，图像超分辨率技术引入了知识蒸馏（Li 等，2021）和注意力机制（周波 等，2021），使重建的高分辨率图像具有更好的视觉效果。

1.2　视频超分辨率

视频超分辨率技术旨在根据已有的低分辨率视频序列生成具有真实细节和内容连续的高分辨率视频序列。定义 $\boldsymbol{I}_l$ 为低分辨率视频序列，$\boldsymbol{I}$ 为原始高分辨率视频序列，低分辨率视频的获取过程可表示为

$$\boldsymbol{I}_l = \varphi(\boldsymbol{I}; \alpha) \tag{1}$$

式中，φ 函数表示原始高分辨率视频退化成低分辨率视频的退化映射函数。α 表示在退化过程中影响该过程的各种参数，如噪声因子、下采样因子和运动模糊因子等。利用视频超分辨率网络重建高分辨率帧的过程可视为式（1）的逆过程，表示为

$$\bar{\boldsymbol{I}} = \varphi^{-1}(\boldsymbol{I}_l; \theta) \tag{2}$$

式中，$\bar{\boldsymbol{I}}$ 表示估计的高分辨率视频，θ 为该过程中的各项训练参数。在视频超分辨率网络的训练过程中，需要对退化过程进行重新建模，一般采用增加模糊核的双线性插值下采样，具体为

$$\boldsymbol{I}_l = (\boldsymbol{I} \otimes k) \downarrow_s + n \tag{3}$$

式中，k 表示模糊核，n 为高斯白噪声，$\otimes$ 为原始高分辨率视频图像与模糊核的卷积运算，$\downarrow_s$ 表示比例为 s 的下采样。视频超分辨率的重建目标为

$$\hat{\theta} = \arg\min_{\theta} \boldsymbol{M}(\bar{\boldsymbol{I}}, \boldsymbol{I}) + \lambda\phi(\theta) \tag{4}$$

式中，$\boldsymbol{M}(\bar{\boldsymbol{I}}, \boldsymbol{I})$ 为高分辨率视频图像的估计值与原始高分辨率视频帧之间的损失。λ 为权衡参数，ϕ 函数为正则化项，$\hat{\theta}$ 为优化后的网络参数。

视频与图像的区别在于视频帧间存在运动信息，如果将视频转换为多个单帧的图像进行超分辨率重建，处理后的视频存在波浪效果以及大量伪影。因此，帧间时空信息的利用对生成的视频质量有决定性影响。该因素也是本文总结基于深度学习的视频超分辨率技术的分类标准。

1.3　评价标准

在视频超分辨率的研究中，评价指标能够有效地判断高分辨率视频的质量，从而对神经网络进行优化。主观评价指标的获取需要大量的精力和财力，难以度量所有模型，且方法太多难以概述。因此，本文主要对客观评价指标进行总结介绍。评价图像质量主要通过峰值信噪比（peak signal-to-noise ratio，PSNR）（Sheikh 等，2006）和结构相似性（structural similarity，SSIM）（Wang 等，2004）这两类指标。

1.3.1　峰值信噪比

PSNR 是信号的最大可能功率与影响信号质量的噪声功率之间的比率。因为许多信号具有非常宽的动态范围，所以 PSNR 通常用对数分贝标度来表示。PSNR 的计算基于均方误差（mean squared error，MSE）。对于彩色视频图像，MSE 分别对每个彩色像素的相似度进行衡量，PSNR 在 MSE 基础上进一步量化图像的质量差距。在视频超分辨率算法比较过程中，常用的数据集都为 8 bit/彩色视频图像，PSNR 普遍能达到 20～40 dB，若能接近 40 dB，则说明重建质量极好。

1.3.2　结构相似性

SSIM 是一种衡量两幅图像相似度的指标。

SSIM的计算基于两图的亮度(均值)、对比度(标准差)和结构相似度(协方差)等3个角度。SSIM值越高,视频帧的重建质量越趋近于真实高分辨率帧,在一定程度上证明视频超分辨率算法的设计越高效。

PSNR和SSIM计算简单,且能够在一定程度上反映视频序列的失真程度,几乎所有视频超分辨率方法都使用二者作为评价指标。

1.3.3 其他评价标准

比较视频数据相似性有很多种直接的数学方法,比如汉明距离、编辑距离以及欧几里得距离等,但是这些方法测量的结果往往与人主观的视觉感受有较大出入。为了更好地对视频超分辨率问题评估重建视频的质量,一些基于视觉感知的指标应用于评价视频超分辨率的生成结果中。NTIA(national telecommunications and information administration)在2001年设计了通用视频质量模型(video quality model,VQM),利用统计学原理模拟实际的人眼视觉系统,提取参考及测试视频中人眼可感知的特征,包括模糊、帧间运动和噪音等信息。Wolf和Pinson(2011)设计了VFD(variable frame delay)算法,将每个接收到的帧和一系列原始视频帧进行逐像素比较,选择最可能匹配的帧,并基于VFD算法,提出一种改进指标VQM-VFD。

通过引入基于人类视觉系统的感知模型,MOVIE(motion-based video integrity evaluation)(Seshadrinathan和Bovik,2010)可进一步提升视频质量评估的准确性。这种方法计算视频中物体的运动矢量、联合时域和空域的失真信息,能提供与人类主观判断更加吻合的质量分数。SOFVSR(video super-resolution through hr optical flow estimation)(Wang等,2018b)等视频超分辨率方法都尝试使用MOVIE来评估网络性能。但是MOVIE的运算复杂度远高于前面提及的几种评价方法,因而使用范围有限。

Zhang等人(2018a)提出基于特征图的LPIPS(learned perceptual image patch similarity)方法,在特征空间评估不同网络中参考视频块和失真视频块的特征距离。该方法利用固定的卷积层,逐层计算网络输出的余弦距离,然后取平均值,从而可以捕获更多的语义相似性。

以上方法都是针对单幅图像进行评估,视频超分辨率任务要保证视频的连续性,这也是区别于单幅图像超分辨率任务的难点之一。像素差异和感知变化对于量化真实的时间一致性至关重要。Teco-Gan(temporal coherence via self-supervision for gan-based video generation)方法(Chu等,2020)中,为了对时间连贯性进行度量,提出两个新的评价指标tOF(tandem optical flow)和tLP(tandem perceptual LPIPS)。tOF利用Farneback算法(Farnebäck,2003)估计的光流计算视频序列的逐像素差异,而tLP则使用LPIPS方法生成的深度特征图来度量视频随时间的感知变化。

1.4 数据集

视频超分辨率算法使用的数据集主要分为YUV(luminance,chroma)格式和RGB格式。本文常用的数据集如表1所示,数据规模为视频数×帧数,由于YUV25数据集和UVG(ultra video group)数据集中不同视频的帧数不完全相同,因此不一一列举,而用"25×-"和"16×-"表示。在此介绍的YUV格式数据集表示的是YCbCr编码的视频文件,Y表示明亮度(即灰度值)。在使用YUV格式的数据集时,研究人员通常在Y通道利用数据对视频超分辨率网络进行训练与测试(Isobe等,2020b;Dong等,2016;Huang等,2015;Wang等,2015a)。

最早使用的YUV高分辨率视频数据集是YUV25。由于数据集分辨率较低,因此常被用于2倍和3倍放大因子的视频超分辨率研究。2014年构建的YUV21数据集在不同场景中捕获了21种具有不同类型动作的视频,但分辨率仅为352×288像素。目前最新的YUV数据集是由芬兰的学术视频编码组Ultra Video Group构建的UVG数据集,分辨率达到3 840×2 160像素。该数据集不断扩充数据,至2020年,已包含16种4 K,120帧/s的视频序列。UVG数据集既有包含丰富人脸特征信息的视频,也有包含大量动作变化的视频。YUV占用带宽小,适合传输和存储。相对而言,RGB视频更适用于显示系统。RGB格式的Vid4(video 4)数据集(Li和Wang,2017)包含城市、日历、步行和树叶4个常见场景的视频。4个视频的帧数分别为34、41、47和49,分辨率分别为704×576像素、720×576像素、720×480像素和720×480像素。Vid4广泛使用于视频超分辨率研究,也是本文实验部分的测试集之一。

此前RGB数据集包含的视频内容不够丰富。Tao等人(2017)为了提升DRVSR(detail-revealing

表 1　视频超分辨率数据集

Table 1　Datasets of video super-resolution

数据集	年份	色彩模式	常用类型	下载链接	数据规模	分辨率/像素
YUV25(Protter等,2009)	–	YUV	训练集	https://media.xiph.org/video/derf/	25×–	–
Vid4(Liu和Sun,2011)	2011	RGB	测试集	https://drive. google. com/drive/folders/10gUO6zBeOpWEamrWKC tSkkUFukB9W5m	49(Foliage) 47(Walk) 41(Calendar) 34(City)	720×480(Foliage) 720×480(Walk) 720×576(Calendar) 704×576(City)
YUV21 (Li和Wang,2017)	2014	YUV	测试集	http://www.codersvoice.com/a/webbase /video/08/152014/130.html	21×100	352×288
Myanmar (Kappeler等,2016)	2014	RGB	训练集	https://www. harmonicinc. com/insights/blog/4k-in-context/	1×527	2 840×2 160
CDVL (Caballero等,2017)	2016	RGB	训练集	http://www.cdvl.org/	100×30	1 920×1 080
SPMCS(Tao等,2017)	2017	RGB	训练集+测试集	https://tinyurl.com/y426dcn9	975×31	540×960
MM522(Wang等,2019c)	2019	RGB	训练集	https://github.com/psych/MMCNN/	542×32	1 280×720
UDM10(Yi等,2019)	2019	RGB	测试集	https://github.com/psychopa4/PFNL	10×32	1 272×720
Vimeo90K(Xue等,2019)	2019	RGB	训练集+测试集	http://toflow.csail.mit.edu/	91 701×7	448×256
REDS(Nah等,2019a)	2019	RGB	训练集+测试集	https://seungjunnah.github.io/Data sets/reds.html	270×100	1 920×1 080
UVG(Ebadi等,2017)	2017	YUV	测试集	http://ultravideo.cs.tut.fi/	16×–	3 840×2 160
RealVSR(Yang等,2021)	2021	RGB	训练集+测试集	https://drive.google.com/drive/folders/1-8MvMEYMOeOE713DjI7TJKyRE-LnrM3Y	200×50	1 024×512

deep video super-resolution)中亚像素运动补偿模块的性能,构建了包含大量运动细节的SPMCS(sub-pixel motion compensation)数据集,该数据集包含975个31帧视频,每帧分辨率为540 × 960像素。PFNL(progressive fusion video super-resolution network)(Yi等,2019)利用MM522(multi-memory 522)数据集进行训练。不同于此前广泛使用的仅包含4个场景的低分辨率视频的测试数据集Vid4,构建了UDM10(ultra dense memory)数据集。UDM10包含10个场景,每帧分辨率为1 272 × 720像素。

2019年,Xue等人(2019)设计ToFlow(video enhancement with task-oriented flow),针对视频插帧、视频去噪、视频解码和视频超分辨率等不同视频处理任务构建对应的光流分析网络。通过剪辑从vimeo. com下载的89 800个视频,构建了一个大规模、高质量的视频数据集Vimeo-90K(Xue等,2019)。此数据集涵盖了大量的动作场景,按照研究的视频任务划分为Triplet数据集和Setuplet数据集。Triplet数据集用于研究视频插帧技术。Setuplet数据集用于研究视频去噪、视频解码和视频超分辨率技术。Setuplet数据集由91 701个7帧视频序列组成,固定分辨率为448 × 256像素。本文介绍的EDVR(video restoration with enhanced deformable)(Wang等,2019b)、TDAN(temporal deformable alignment network)(Tian等,2020)、DNLN(deformable non-local network)(Wang等,2019a)和MEMC-Net(motion estimation and motion compensation driven neural network)(Bao等,2019b)等视频超分辨率网络都使用Vimeo-90K数据集进行训练测试。

NTIRE19(new trends in image restoration and enhancement)挑战赛(Nah等,2019b)中,举办方提供了包含大量真实动态场景的REDS(realistic and

dynamic scenes)数据集。该数据集由300个分辨率为720×1 280像素的100帧视频序列组成。其中，训练集、验证集和测试集数量分别为240、30、30。由于此次大赛提出大量出色的视频超分辨率方法，许多学者依据REDS数据集与这些前沿方法进行性能比较。

Yang等人(2021)构建了首个移动端真实场景的视频超分辨率数据集，使用iPhone 11 Pro Max上两个不同焦距的相机和DoubleTake软件拍摄不同尺度的成对LR-HR(low resolution-high resolution)视频序列，52 mm镜头采集的数据作为HR序列，26 mm镜头采集的数据作为LR序列，最终挑选500对LR-HR序列，每对序列包含50帧分辨率为1 024×512像素的视频。相比传统下采样方法生成数据训练的模型，在RealVSR(real video super-resolution)上训练的模型具有更真实的细节与丰富的纹理。未来会有更多针对不同运动场景的真实视频超分辨率数据集，进一步优化模型训练，提升重建视频的质量。

2 基于图像配准的视频超分辨率方法

由于不同视频帧之间存在运动差异，直接将多帧图像融合重建会导致生成的视频出现边缘模糊、细节混淆等影响视觉效果的现象。因此，研究人员往往先对相邻帧与目标帧进行配准。

传统的配准方法包括过平移、翻转和旋转等预处理操作，而这些方式包含复杂的计算过程且达不到好的配准效果。现在基于深度学习的视频帧配准方法可分为基于光流的配准方法、基于卷积核估计的配准方法和基于可变形卷积的配准方法3种。

2.1 基于光流的配准方法

为了充分利用帧间时空信息，研究人员提出了各种方法进行运动估计与运动补偿，将相邻帧与目标帧对齐，然后在此基础上融合特征，重建高分辨率视频。Lertrattanapanich和Bose(1999)设计投影模型来模拟相机的运动参数，利用相机运动来补偿帧间运动。张义轮等人(2013)通过光流的初始运动估计和精细的块匹配，利用相似信息不断修正迭代反投影中的视频重建误差。Liu和Sun(2014)使用贝叶斯自适应方法，以最大后验(maximum a posterior, MAP)的方式迭代估计所有算子。传统的基于插值的超分辨率方法已证明运动补偿对于高分辨率视频重建的有效性。而在视频超分辨率领域，光流法是最常用的运动估计与运动补偿方法。1950年首次提出了光流的概念。光流是空间运动物体在观测成像面上像素运动的瞬时速度。如图3所示，利用二维图像上特定坐标点灰度瞬时变化率来定义物体的3D运动效果，灰度瞬时变化率也称为光流矢量。

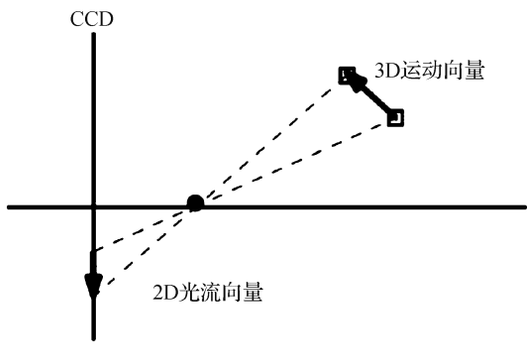

图3 三维运动在二维平面上的投影

Fig. 3 Projection of 3D motion on a 2D plane

基于光流法进行运动估计时，存在两个主要的假设。一是亮度恒定。随着时间的改变，相邻帧之间的同一像素目标运动时，其亮度(灰度值)不会发生改变。所有的光流法都基于此假设条件。二是小运动。相邻帧之间同一像素的位移不可过大，即随着时间的变化，像素点不会产生剧烈的位置变化。这是可以利用灰度变化进行光流计算的原因，亦是所有光流法都不可或缺的假设条件。

考虑视频中某一个像素$\boldsymbol{I}(x,y,t)$在第1帧的光强(t代表其所在的时间维度)下用了$\mathrm{d}t$时间移动了$(\mathrm{d}x,\mathrm{d}y)$的距离到下一帧。依据上文提到的亮度恒定假设，可得

$$\boldsymbol{I}\left(x,y,t\right)=\boldsymbol{I}\left(x+\mathrm{d}x,y+\mathrm{d}y,t+\mathrm{d}t\right) \tag{5}$$

将式(5)右端进行泰勒展开，可得

$$\boldsymbol{I}\left(x,y,t\right)=\boldsymbol{I}\left(x,y,t\right)+\frac{\partial \boldsymbol{I}}{\partial x}\mathrm{d}x+\frac{\partial \boldsymbol{I}}{\partial y}\mathrm{d}y+\frac{\partial \boldsymbol{I}}{\partial t}\mathrm{d}t+\varepsilon \tag{6}$$

式中，ε代表二阶无穷小项，可忽略不计。将式(6)代入式(5)，可得

$$\frac{\partial \boldsymbol{I}}{\partial x}\mathrm{d}x+\frac{\partial \boldsymbol{I}}{\partial y}\mathrm{d}y+\frac{\partial \boldsymbol{I}}{\partial t}\mathrm{d}t=0 \tag{7}$$

假设$\boldsymbol{u}$和$\boldsymbol{v}$分别为光流沿X轴与Y轴的速度矢量，可得

$$\boldsymbol{u}=\frac{\mathrm{d}x}{\mathrm{d}t},\ \boldsymbol{v}=\frac{\mathrm{d}y}{\mathrm{d}t} \tag{8}$$

令 $\boldsymbol{I}_x=\frac{\partial \boldsymbol{I}}{\partial x}, \boldsymbol{I}_y=\frac{\partial \boldsymbol{I}}{\partial y}, \boldsymbol{I}_t=\frac{\partial \boldsymbol{I}}{\partial t}$ 分别表示图像中像素点的灰度沿 x, y, t 方向的偏导数。综上，式(7)可以写为

$$\boldsymbol{I}_x u+\boldsymbol{I}_y v+\boldsymbol{I}_t=0 \tag{9}$$

式中，$\boldsymbol{I}_x, \boldsymbol{I}_y, \boldsymbol{I}_t$ 均可由图像数据求得，$(\boldsymbol{u}, \boldsymbol{v})$ 即所求光流矢量。

此时，存在一个式子中含有多个未知数的问题。为此，在求光流时，Lucas 和 Kanade(1981)提出 Lucas-Kanade(L-K)光流法。该方法提出了新的假设空间一致性，即前一视频帧中相邻像素点在后一视频帧中也是相邻的。当取一个包含多个像素点的像素块代入原式中，可以获得像素点数目的方程，并利用最小二乘法进行求解，这是当前使用最为广泛的光流估计算法。此外，Druleas 算法(Drulea 和 Nedevschi，2011)也是较为常用的传统方法。Cui 等人(2014)、Wang 等人(2015b)、Cheng 等人(2012)也提出了估计密集的光流场的方法。

由于高分辨率视频成像平面上像素运动的瞬时速度较大，无法满足光流计算中的小运动假设，因此很难直接利用传统方法进行光流计算。于是一些方法提出使用金字塔模型，首先对高分辨率视频帧进行下采样，通过降低图像尺寸来计算粗糙的光流，将相邻帧进行对齐后，再不断向上采样获得更为精细的光流，最终计算出准确的光流信息。近年来，上述过程通常利用深度卷积神经网络来实现。基于深度学习的光流网络包括 FlowNet(Dosovitskiy 等，2015)、FlowNet2.0(Ilg 等，2017)、EpicFlow(edge-preserving interpolation of correspondences flow)(Revaud 等，2015)、SpyNet(spatial pyramid network for optical flow)(Ranjan 和 Black，2017)和 PwcNet(using pyramid，warping，and cost volume for optical flow)(Sun 等，2018a)等。到目前为此，光流配准的视频超分辨率方法数量占比最多，本文将其细粒度地划分为基于常规卷积的方法、基于循环神经网络的方法、基于 GAN 的方法和基于 Transformer 的方法，并进行详细阐述。

2.1.1　基于常规卷积的光流配准方法

2011 年，Shahar 等人(2011)设计了基于光流的传统视频超分辨率方法。随着深度学习技术不断发展，Liao 等人(2015)首次将卷积神经网络运用于视频超分辨率的研究，设计了深度候选图集成学习方法(deep draft-ensemble learning method，Deep-DE)，利用 TV-$l1$ 光流(Brox 等，2004)和运动细节保留光流(Xu 等，2012)，通过调节两者权重，生成不同正则化的运动估计结果，然后利用双线性插值法获得生成超分辨率候选图。最终，将所有超分辨率候选图像放入卷积神经网络中进行融合。基于常规卷积的光流配准方法由此开始发展，此类方法的总结如表 2 所示。

表 2　基于 2D 卷积的光流配准方法

Table 2　2D convolution methods with optical flow alignment

方法	主要思想	放大倍率
Deep-DE(Liao 等，2015)	CNN，高分辨率候选图	×4
VSRnet(Kappeler 等，2016)	自适应运动补偿	×2，×3，×4
RVSR(Liu 等，2017)	整流光流法	×4
DRVSR(Tao 等，2017)	亚像素运动补偿	×2，×3，×4
VESPCN(Caballero 等，2017)	实时性，亚像素卷积，空间变换网络	×3，×4
STTN(Kim 等，2018)	时空变换网络	×4
SOFVSR(Wang 等，2018a)	HR 光流估计	×4
MEMC-Net(Bao 等，2019b)	上下文提取网络	×4
MultiBoot VSR(Kalarot 和 Porikli，2019)	FlowNet 2.0	×4
TOFlow(Xue 等，2019)	面向任务的光流网络	×4

Deep-DE 为基于光流法的视频超分辨率技术提供了基础的思路。该网络首先利用光流法对相邻帧与目标帧进行运动估计，然后运用运动补偿对齐相邻帧，提取目标帧与相邻帧特征之后，可以利用逆卷积等方法进行特征融合，以获得高分辨率视频。Kappeler 等人(2016)设计了 VSRnet(video super-

resolution with convolutional neural networks)，提出通过联合处理多个输入帧来提高重建视频质量的思路。在该思路上，选择Druleas算法(Drulea和Nedevschi，2011)进行运动估计。即使帧间存在大位移的像素点，该模块依靠CLG(combined local-global)变分法可以粗略估计光流。然后，使用自适应运动补偿方法(adaptive motion compensation，AMC)降低重建过程中未对齐的相邻帧的影响。VSRnet(video super-resolution net)的FSE(filter symmetry enforcement)模块将卷积核反向传播过程中的梯度同时应用于对称的卷积核，以此减少参数计算并加速训练，也证实了此种方法的高效性。

上述两种方法将光流对齐的低分辨率视频帧联合传递到卷积神经网络，以重建每个高分辨率视频帧。然而这些方法中的运动估计、运动补偿模块与神经网络分离，很难获得整体的最优解。

为了满足及时性的需求，部分学者提出将视频超分辨率问题转换为多个单帧的图像超分辨率问题的思路，但明显忽略了帧之间的时序信息，生成的视频图像中存在大量伪影，进而影响视频连贯性。高效视频亚像素卷积网络(video efficient sub-pixel convolutional network，VESPCN)是Caballero等人(2017)设计的第1个端到端训练的视频超分辨率网络，一些方法提出同时结合多个连续视频帧之间的时空信息进行研究。如图4所示，该网络由运动估计模块、运动补偿模块和多帧融合模块组成。图中$\boldsymbol{I}_t^{\mathrm{LR}}$表示目标时刻低分辨率帧，$\boldsymbol{I}_{t-1}^{\mathrm{LR}}$和$\boldsymbol{I}_{t+1}^{\mathrm{LR}}$表示前一时刻和后一时刻的低分辨率参考帧，$\boldsymbol{I}_t^{\mathrm{SR}}$表示目标帧的超分辨率结果，下同。

运动估计模块利用金字塔网络解决帧间可能存在大位移像素点的问题，实现从粗略到精细的光流估计过程。运动补偿模块利用了空间变换网络(spatial transformer networks，STN)(Jaderberg等，2016)。STN可以推断两图像之间空间映射参数，且已成功用于无监督训练光流特征编码(Ahmadi和Patras，2016；Ganin等，2016；Handa等，2016；Patraucean等，2016)。VESPCN首次将该模块引入视频帧的配准过程，STN成为常用的运动补偿方法。

时空亚像素卷积模块将以目标帧为对称中心的连续输入帧融合。针对VSRnet需要双三次上采样，导致网络的计算效率大幅降低的问题，ESPCN(Shi等，2016)使用实时图像超分辨率方法进行上采样。

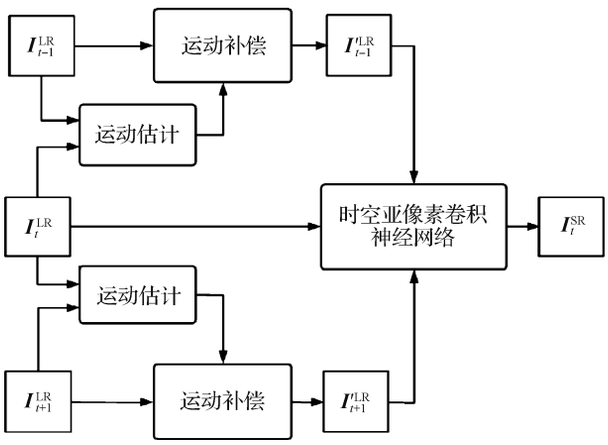

图4 VESPCN的网络结构(Caballero等，2017)

Fig. 4 The network architecture of VESPCN (Caballero et al.，2017)

通过亚像素卷积(sub-pixel convolution)这种高效、快速且无参的像素重排列的方式上采样可以获得高分辨率视频帧。由于VESPCN在速度和性能上的优势，后续很多方法都是基于该网络框架改进运动估计模块、运动补偿模块以及超分辨率重建模块。

Liu等人(2017)提出的RVSR(robust video super-resolution)由空间对齐模块和时间自适应模块组成，网络结构如图5和图6所示。空间对齐模块运用整流光流法(rectified optical flow)(Liu等，2017)预测相邻帧到目标帧的空间变换参数，将像素级别的运动数值约为整数，以避免使用可能导致模糊或混叠的插值算法。依靠回归后得到的水平位移和垂直位移两个变换参数，空间对齐模块可实现低分辨率相邻帧的对齐。时间自适应网络包含多条超分辨率推断分支、一条时间调制分支和时间聚合网络。每条超分辨率推断分支利用ESPCN将连续的视频帧进行上采样获得高分辨率图像。时间调制分支获取每个高分辨率视频帧的权重图。超分辨率推断分支生成的高分辨率图像和时间调制分支的权重点乘后，利用时间聚合网络融合为高分辨率视频序列。时间自适应模块的损失是真实数据与网络生成的高分辨率视频之间的误差。空间对准模块的损失是利用整流光流对准获得的变换参数与其他定位网络估计的变换参数之间的误差。RVSR(Liu等，2017)组合两模块的损失函数，采用端到端的联合学习方式训练模型。

在VESPCN的基础上，显示细节的深度视频超分辨率网络(detail revealing deep video super- resolu-

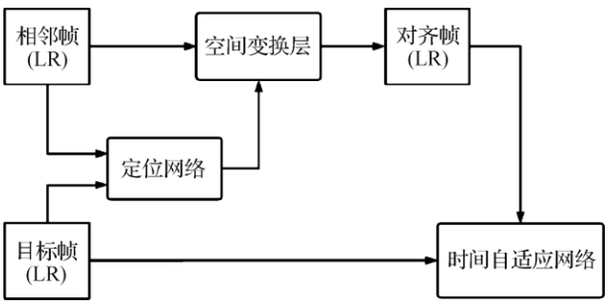

图5　空间对齐网络架构(Liu等,2017)

Fig. 5　The network architecture of spatial alignment module (Liu et al. ,2017)

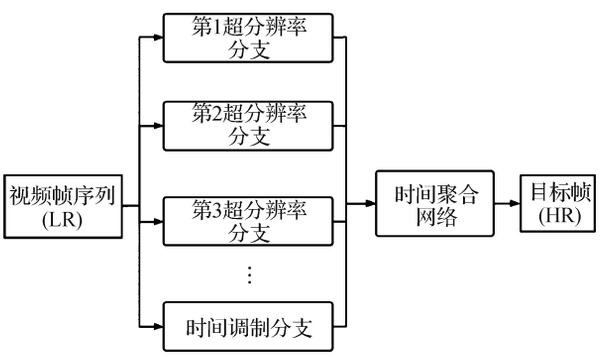

图6　时间自适应网络架构(Liu等,2017)

Fig. 6　The network architecture of temporal adaptive module (Liu et al. ,2017)

tion,DRVSR)(Tao等,2017)提出新的运动补偿结构。如图7所示,设计了由网格生成器和采样器组成的亚像素运动补偿模块(sub-pixel motion compensation,SPMC)。SPMC将光流配准与亚像素卷积相结合,在相邻帧运动补偿过程中,同时上采样直接生成高分辨率对齐帧。DRVSR设计了编码器—解码器(Mao等,2016)结构的细节融合模块,并在模块中添加ConvLSTM(Shi等,2015)网络处理时序信息,同时,网络大量利用跳跃连接以加强网络内部的信息流动性。

上述网络每次都只能估计两个连续帧(相邻帧和目标帧)之间的光流,每个时间步长需要计算N次光流,以处理N个目标帧。针对上述算法的低效,Kim等人(2018)在STN基础上设计了时空变换网络(spatial-temporal transformer network,STTN),STTN不仅在空间域而且在时间域中也可以改变特征图。网络结构如图8所示,STTN的时空光流估计模块可以代替之前的方法,用较少的计算量在单个时间步长内同时处理多个视频帧。运动估计后,时空光流估计模块可以输出代表帧之间时空变化的三维流。计算出的时空流与经过时空采样器模块插值的视频帧融合后,STTN利用超分辨率模块重建特征图,一次性地获得多目标帧的高分辨率序列。

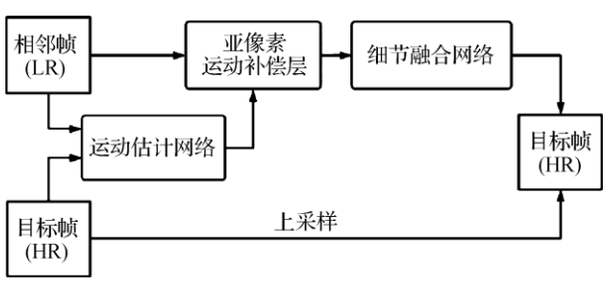

图7　DRVSR的网络架构(Tao等,2017)

Fig. 7　The network architecture of DRVSR(Tao et al. ,2017)

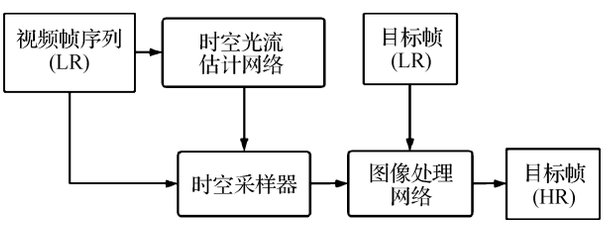

图8　STTN的网络结构(Kim等,2018)

Fig. 8　The network architecture of STTN(Kim et al. ,2018)

传统的视频超分辨率方法(Chen等,2018;Kim等,2019b)已经证明将图像和光流同时进行超分辨率处理会产生更高质量的视频。高分辨率光流比低分辨率光流能提供更精确的对应关系,以此提高视频帧间配准结果的精度。Wang等人(2018a)研究的高分辨率光流视频超分辨率方法(super-resolution optical flow for video super-resolution,SOFVSR)利用OFRNet(optical flow reconstruction network)直接获取高分辨率光流进行运动补偿,如图9所示。按照传统方法(Bouguet,2001),采用3级金字塔模型来处理复杂的运动模式。在前两层金字塔中,SOFVSR利用卷积以及残差密集块获取两倍下采样的低分辨率视频帧的光流。第3层金字塔网络与前两层有相似复杂的运动模式,但是最后的光流估计网络被替换为亚像素卷积,直接上采样低分辨率光流生成高分辨率光流。SOFVSR采用空间到深度转换网络(space-to-depth transformation)(Sajjadi等,2018)将高分辨率光流映射到低分辨率图像网格。然后,OFRNet的输出进入SRNet(super-resolution network)中,经过多个卷积层,连续的残差密集块和亚像素卷积生成高分辨率视频。结合OFRNet和SRNet的损

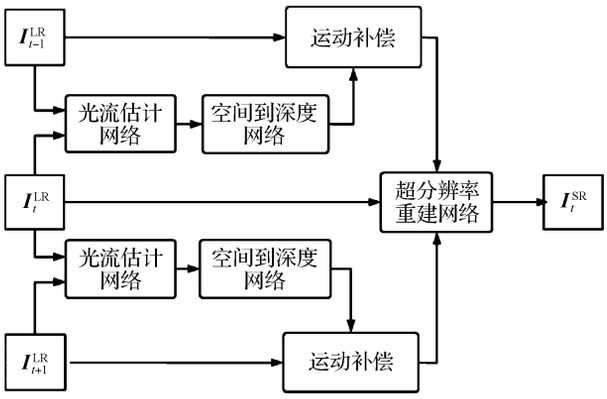

图9　SOFVSR的网络结构(Wang等,2018a)

Fig. 9　The network architecture of SOFVSR (Wang et al. ,2018a)

失函数训练模型,最终获得很好的重建效果。

随着光流网络的发展,很多方法尝试将新的光流网络引入视频超分辨率问题的研究。其中,多阶段多参考自助法(multi-stage multi-reference bootstrapping method, MultiBoot VSR)(Kalarot和Porikli, 2019)在输入子网中利用FlowNet 2.0进行运动估计。FlowNet 2.0相较于FlowNet 1.0在网络结构上进行改进。FlowNet 2.0将FlowNetCorr网络和两个FlowNetSimple网络进行堆叠,同时构建多个分支网络。经过以上策略调整后,推断速度较之前有一定的下降,但错误率在原来基础上降低了50%以上。

各个输入子网生成的低分辨率特征图合并后作为混合主干网络的输入。混合主干网络由多个全卷积的残差块组成,并具有从第1个残差块到最后一个残差块的跳越连接。最后利用空间上采样网络进行高分辨率视频的重建。

在对光流估计网络的改进过程中,Xue等人(2019)针对视频超分辨率、视频去噪和视频插帧等不同任务研究不同的光流学习网络,设计了任务导向型光流网络(task-oriented flow, ToFlow)。由于光流估计模块与网络的其余部分联合训练,因此学习到特定任务中最适合表达特征的光流。图10展示的是视频超分辨率任务的网络结构,利用FlowNet将相邻的视频帧利用3层金字塔模型由粗到细地估计光流,并将光流图像与相邻帧放入结合双线性插值法的空间变换网络进行对齐操作,最终将连续的视频帧特征放入图像处理模块生成高分辨率视频。

基于常规卷积的光流配准方法发展最久,形成了基于滑动窗口法的视频超分辨率基本框架。对齐、融合和重建成为所有图像配准方法的基本流程,方法设计过程中构建了包括Vimeo90K和REDS在内的优秀数据集。但是单一2D卷积难以充分处理视频时序信息,视频重建效果达到瓶颈,基于其他算法的视频超分辨率网络也逐渐提出。基于常规卷积的光流配准方法对运动估计模块的设计极具参考价值,成为很多后续工作的参照网络(baseline)。

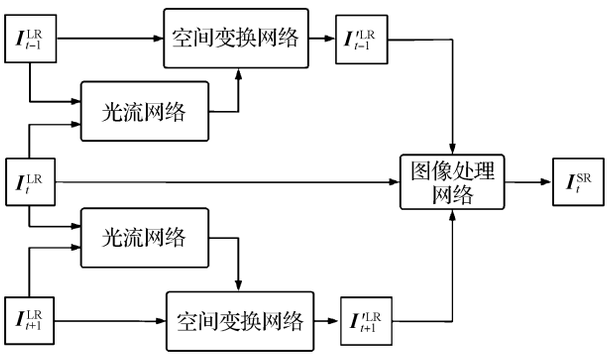

图10　ToFlow的网络结构(Xue等,2019)

Fig. 10　The network architecture of ToFlow (Xue et al. ,2019)

2.1.2　基于循环神经网络的光流配准方法

常规卷积的光流配准方法大多利用滑动窗口法同时处理多帧,以重建高分辨率目标帧。滑动窗口的感受野越大,同时处理的帧越多,目标帧的重建质量越好。远距离帧中的信息有利于细节恢复,通病是在每个视频的两端PSNR的差异很大。循环神经网络则能在一定程度改善边缘视频帧的视觉质量,并且可以更快速地同时处理低分辨率帧以获取高分辨视频序列,表3总结了基于循环神经网络的光流配准方法。

表3　基于循环神经网络的光流配准方法

Table 3　RNN methods with optical flow alignment

方法	主要思想	放大倍率
RRCN(Li等,2019a)	CLG-TV光流法	×2,×3,×4
MMCNN(Wang等,2019c)	多记忆细节融合网络	×2,×3,×4
RBPN(Haris等,2019)	多重投影	×4
BasicVSR(Chan等,2021a)	双向传播,光流对齐	×4
IconVSR(Chan等,2021a)	信息填充机制	×4
BasicVSR++(Chan等,2021b)	光流指导可变形卷积	×4

RRCN(residual recurrent convolutional network)(Li等,2019a)是第1个用于视频超分辨率的深度循环神经网络。如图11所示,RRCN由前馈网络与后

馈网络两部分组成，RRCN 利用 CLG（combined local-global）变分光流法进行运动估计，并将经过光流对齐的帧与目标帧作为输入，相应获取正向残差和反向残差。且 RRCN 最终会融合所有输出的残差图，以还原目标帧中的微小细节。考虑到使用数据增广对模型训练的优化，可使用自集成（self-ensemble）的方式旋转和翻转输入视频帧，以生成多个不同的输入组，然后将它们分别馈送到超分辨率网络，并计算所有输出图像中逐像素的平均值，获取最终的高分辨率视频。自集成方法的 RRCN+与 EDSR+拼接在一起可获得效果显著的算法 RRCN++。

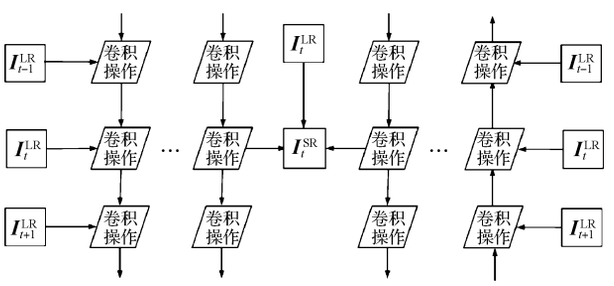

图 11　RRCN 的网络结构（Li 等，2019a）

Fig. 11　The network architecture of RRCN（Li et al.，2019a）

常规卷积的光流配准方法的运动估计、运动补偿模块的高效性已得到证明，许多基于循环神经网络的光流配准方法大多直接在这些网络基础上改进特征融合模块和高分辨率视频重建模块。Wang 等人（2019b）提出的多记忆卷积神经网络（multi-memory convolutional neural network，MMCNN）直接利用 VESPCN 中的 MCT（spatial transformer motion compensation）模块配准视频帧，在帧间信息融合模块引入循环神经网络，设计了多记忆细节融合模块。MMCNN 的网络结构如图 12 所示，多记忆细节融合模块利用 ConvLSTM 处理相邻帧与目标帧间的时序信息。ConvLSTM 的结构将在 4.3 节进行详细介绍。多记忆细节融合模块包含大量的密集残差网络，以此减少参数并防止梯度消失。最后将所有特征图放入亚像素卷积层，与双三次插值上采样的目标帧相结合得到高分辨率视频帧。

基于光流法的视频超分辨率研究越来越多，很多方法开始在网络结构上进一步创新。受反投影网络（Yi 等，2019；Bao 等，2019a；Lai 等，2017）的影响，递归反投影网络（recurrent back-projection network，RBPN）（Haris 等，2019）在 RNN 基础上利用了编码—

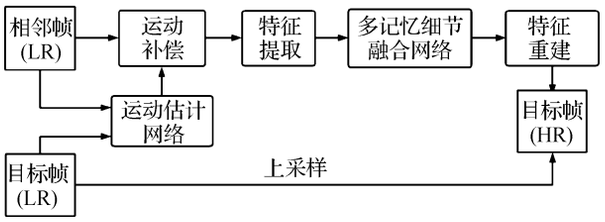

图 12　MMCNN 的网络结构（Wang 等，2019c）

Fig. 12　The network architecture of MMCNN（Wang et al.，2019c）

解码（encoder-decoder）模型。如图 13 所示，初始特征提取模块设置了多图像超分辨率（multi images super-resolution，MISR）通道和单图像超分辨率（single image super-resolution，SISR）通道。MISR 将目标帧、相邻帧和计算的光流合并，卷积获得多通道的特征张量 M。SISR 直接获得目标帧的特征张量 L。循环网络的多投影模块将第 K 个 M 张量与第 $K-1$ 个 L 张量输入到编码器中上采样。生成的高分辨率特征张量输入解码器中，下采样至低分辨率，并作为下一多投影模块的输入。RBPN 最终卷积多个合并的高分辨率特征，生成高分辨率视频帧。

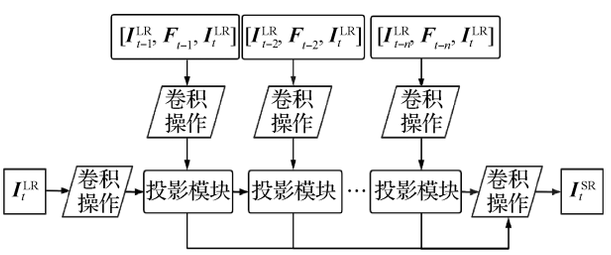

图 13　RBPN 的网络结构（Haris 等，2019）

Fig. 13　The network architecture of RBPN（Haris et al.，2019）

最初基于光流的图像配准方法在利用循环神经网络时，难以发挥 RRN 的优势。主要原因是这些方法都使用光流在图像级别对齐相邻帧，并且都使用多张对齐帧帮助网络重建目标帧，无法摆脱滑动窗口法的限制。BasicVSR（Chan 等，2021a）使用光流对齐相邻帧的特征，以避免图像级别对齐带来的模糊和不准确性。前文也曾介绍，长序列视频累积互补信息有利于重建目标帧。一些方法使用双向循环网络处理使特征实时、独立地向后序隐藏层移动。网络结构如图 14 所示，BasicVSR 利用光流对齐相邻隐藏层的特征信息，最终获得很好的重建效果。

在 BasicVSR 的基础上，进一步引入信息再填充（information-refill）机制和耦合传播（coupled propaga-

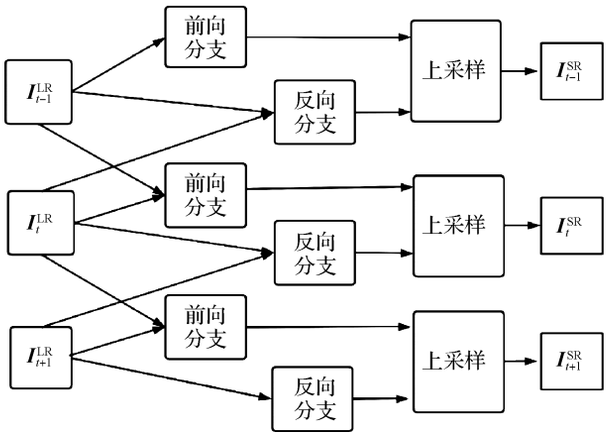

图14　BasicVSR的网络结构(Chan等,2021a)

Fig. 14　The network architecture of BasicVSR (Chan et al. ,2021a)

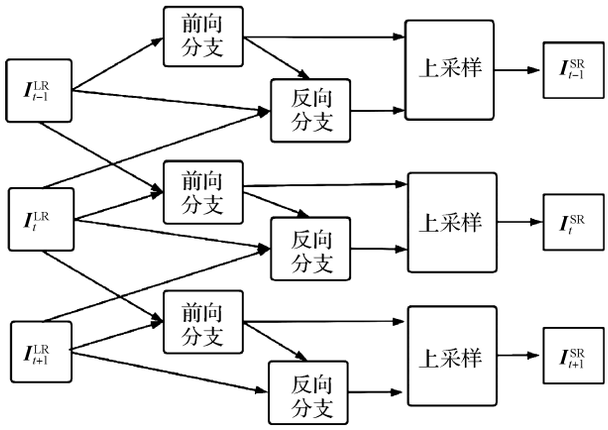

图15　IconVSR的网络结构(Chan等,2021a)

Fig. 15　The network architecture of IconVSR (Chan et al. ,2021a)

tion)机制,进而设计了IconVSR(information-refill mechanism and coupled propagation video super-resolution)(Chan等,2021a)。IconVSR的网络结构如图15所示,耦合传播建立了前向网络与后向网络之间的连接,因此,前向传播分支可以从之前帧和后续帧中获取信息,重建质量更高的视频序列。信息再填充机制指收集关键帧索引,并建立特征提取器,辅助信息传播过程中的关键帧重建,以减少错误累计。相比BaiscVSR,IconVSR能够恢复更精细的细节。在NTIRE2021(new trends in image restoration and enhancement)挑战赛中,Chan等人(2021a)进一步改进BasicVSR,提出的BasicVSR++在视频超分辨率与压缩视频增强竞赛等4个赛道上获取了三冠一亚的成绩,其在RNN基础上进一步设计了网格传播机制,实现交替方式使帧间特征进行双向传播,这样重复地利用视频帧的信息,可以更好地集成特征,但其并不是简单地利用光流法进行运动估计与运动补偿,而是将SpyNet(spatial pyramid network)(Ranjan和Black,2017)获取的光流作为可变形卷积的指导信息,这在2.2节基于可变形卷积的视频超分辨率方法中将进行更详细地描述。

在循环神经网络中利用光流信息来配准相邻图像特征获得了很好的性能,仅需利用极小的参数量即可实现极佳的性能,但由于光流配准误差的存在,运动估计与运动补偿的过程不可能完全正确。在循环神经网络对信息的传递过程中,可能积累了光流估计的误差,进而影响了视频重建帧视觉质量。因此增加对光流配准错误的鲁棒性是值得进一步研究的设计方向。

2.1.3　基于GAN的光流配准方法

加拿大蒙特利尔大学的Goodfellow等人(2014)提出一种生成式对抗网络(GAN)。此后,GAN受到深度学习领域的广泛关注。生成式对抗网络由生成器(generator)与判别器(discriminator)组成,相互博弈学习的过程使两个模型的性能同时得到增强,最后的博弈结果是判别器无法区分真实数据与生成器生成的结果。GAN已用于多种生成任务,如3D建模和音频合成等。在计算机视觉领域,GAN的生成能力最先用于重建高质量的图像。Ledig等人(2017)率先使用GAN网络,结合特征损失生成了极具真实感、高度细致的高分辨率图像。对抗训练在单幅图像超分辨率任务中取得的成功促使越来越多的方法尝试将GAN用于视频超分辨率技术中。

在开始阶段,主要使用常规的GAN损失函数,具体为

$$\min_{\theta}\max_{\phi} L_{\mathrm{GAN}}(\phi,\theta)=E_X\left[\log D_\phi(\boldsymbol{X})\right]+E_Y\left[\log\left(1-D_\phi\left(G_\theta(\boldsymbol{Y})\right)\right)\right] \tag{10}$$

式中,$\boldsymbol{X}$是维度$N\times N$的高分辨率目标帧,$\boldsymbol{Y}$是低分辨率输入帧序列,D_ϕ是可训练参数ϕ的鉴别器,G_θ是具有可训练参数θ的生成网络。这些参数对都应用于可学习网络的卷积核。问题的关键在于仅学习对抗损失提高单帧的质量,生成器网络生成的视频会存在类似于振铃的高频伪影。

德国慕尼黑工业大学的研究人员设计了首个时空判别器,在此基础上提出TecoGAN(temporally

coherent GAN)(Chu 等,2020)。图 16 展示了利用前一高分辨率视频帧的循环生成器,图 16 中"⊕"表示特征图的逐元素相加,下同。在高分辨率视频生成后,时空判别器可以监督高分辨率视频帧的细节信息和时间关系。提出了新的评估指标 tOF 和 tLP,基于动态估计和感知距离来量化视频帧之间的时间连贯度。TecoGAN 利用时空对抗损失 Ping-Pong loss 以提升高分辨率视频帧的准确率和视觉质量,可以有效移除视频内的时间伪影。随着时空判别器的不断增强,以及时空对抗损失的不断下降,循环生成器最终可以保障视频帧之间的连贯性,并生成具有高度真实细节的视频帧。

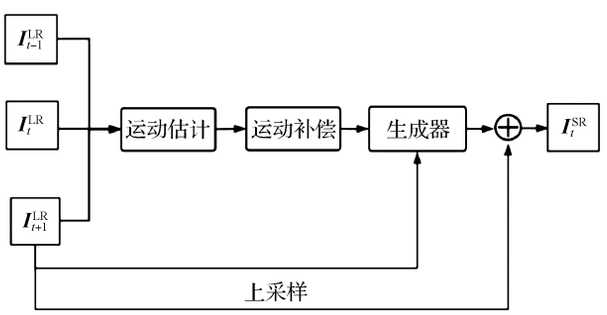

图 16　具备动态补偿的循环生成器(Chu 等,2020)

Fig. 16　The network architecture of the frame recurrent generator with motion compensation(Chu et al. ,2020)

表 4 展示了基于 GAN 的光流配准方法的主要思想和放大倍率。可以看出,基于 GAN 的视频超分辨率方法相对较少,而利用 GAN 进行单幅图像超分辨率的研究工作却非常多。因为鉴别器需要判断视频帧之间的连续性,因而难度较大。但是,如果将视频超分辨率扩充到半监督以及无监督领域时,GAN 将成为重要手段。

表 4　基于 GAN 的光流配准方法

Table 4　GAN methods with optical flow alignment

方法	主要思想	放大倍率
TecoGAN(Chu 等,2020)	GAN,Ping-Pong loss	×4

2.1.4　基于 Transformer 的光流配准方法

Transformer(Vaswani 等,2017)近期获得了广泛的关注,席卷了 NLP(neuro-linguistic programming)和 CV(computer vision)领域,在基于序列的任务上表现出色,谷歌基于 Transformer 推出的 BERT(bidirectional encoder representation from transformers)模型在 11 项 NLP 任务中夺得 SOTA(state-of-the-art)结果。主要是因为与 RNN 和 LSTM 相比,没有递归性,具有并行计算能力和对输入序列长程依赖性的建模能力。在计算机视觉中,Transformer 的优势在于处理视频任务中的时空维度信息。但由于使用时需要高内存和昂贵计算资源,在视频超分辨率领域的应用却非常少。

表 5 展示了基于 Transformer 的光流配准方法的主要思想和放大倍率。可以看出,典型工作为瑞士苏黎世大学的研究人员(Cao 等,2021)提出的 VSR Transformer(video super-resolution Transformer)。Transformer 依赖于线性层计算注意力图,而全连接自注意力层难以挖掘视频数据的时空关系。此外,Transformer 使用的 token 级前馈层缺乏特征对齐能力,处理 token 时忽略了利用不同帧之间的相关性。VSR Transformer 针对上述问题进行了改进,将全连接层替换成卷积神经网络,从而将每一帧的局部信息嵌入所有 token 中。VSR Transformer 在前馈层中添加了双向光流配准特征,将配准特征融合生成细节更精细和边缘更清晰的 HR 帧。

减轻高内存负担和计算资源需求,是 Transformer 应用于视频超分辨率领域的重要挑战,尽管已经提出了几种压缩 Transformer 的方法,但是处理视频 token 和重建更高分辨率视频序列时仍难以胜任。Transformer 的潜力尚未得到充分的利用,相信未来高效的 Transformer 模型能实现视频超分辨率的性能飞跃。

表 5　基于 Transformer 的光流配准方法

Table 5　Transformer methods with optical flow alignment

方法	主要思想	放大倍率
VSR Transformer(Cao 等,2021)	Transformer	×4

2.1.5　基于光流法配准图像的小结

光流估计是视频超分辨率研究中最常用的技术之一,光流网络经过多年研究,架构成熟且效果稳健。但光流法对于复杂的运动难以建模,使用错误的运动估计方法会在基于常规光流对齐后的视频帧中引入伪影,这些伪影会传播到后续的高分辨率视频重建步骤,并影响视频质量,难以保证生成的高分辨率视频的时间连续性。如何更为精准地估计光流以及设计效果更加出色的运动补偿方法,是未来的

重要研究方向。

2.2 基于卷积核估计的图像配准方法

上文所述方法的性能取决于估计的光流的质量。当相邻帧之间存在遮挡、模糊等问题时，计算的光流则会不准确，错误累计进而导致视频超分辨率网络性能低下。

在视频插帧领域，某些研究提出利用动态卷积核进行运动估计与运动补偿。不同于像素级的光流方法，多核集成不仅计算高效，而且具有更强的特征表达能力。Niklaus等人（2017b）提出的AdaConv（adaptive convolution）模型可以估计每个输出像素的空间自适应卷积核。为了减少内存需求，SepConv（separable convolution）方法（Tao等，2017）假设卷积核是可分离的，并使用两个一维卷积核（一个垂直核和一个水平核）来近似2D卷积核，以进一步减少参数量，提升运行速率。

Bare等人（2019）将卷积核估计实现运动估计与运动补偿的方法引入视频超分辨率的研究，设计了实时视频超分辨率方法（real-time video super-resolution，RTVSR），网络结构如图17所示，引用一种编码器—解码器样式的全卷积神经网络作为运动卷积核估计模块，用来估计一对代表水平方向和垂直方向的一维运动卷积核。GEU（gated enhance unit）（Li等，2018）对残差块设置门控单元来改善模型的性能，利用跳跃连接合并GEU内部加权总和的短期信息和GEU输出的长期信息，网络可以学习低分辨率视频帧与高分辨率视频帧间的映射。

AdaConv和SepConv方法进行运动估计时，都不能处理大于预定义内核大小的运动。一些方法尝试联合多种方法来提升视频帧对齐的效果。Bao等人（2019b）将光流法与卷积核估计相结合，提出运动估计和运动补偿网络（motion estimation and motion compensation network，MEMC-Net）。该网络可以应用于视频插帧、视频超分辨率、视频降噪和视频解码等多个方面。如图18所示，在进行视频超分辨率任务时，该网络包含5个部分，即光流估计网络、卷积核估计网络、上下文提取网络、自适应变换网络和帧增强网络。MEMC-Net利用FlowNetS模型估计相邻帧与中心帧的光流，同时将相邻帧与目标帧放入核估计网络，即U-Net（Ronneberger等，2015）后，可获得空间变换的核参数。其中，使用预训练的ResNet18作为上下文提取网络获取目标帧与相邻帧之间的上下文信息。自适应变换网络通过光流和插值内核实现上下文映射和相邻帧的配准。通过额外引入后处理网络，利用配准后的上下文信息消除不正确的流量估计和图像遮挡引起的伪影。MEMC-Net弥补了卷积核估计的运动范围较小和光流法的对齐精度不足的缺点，但带来了庞大的存储压力。

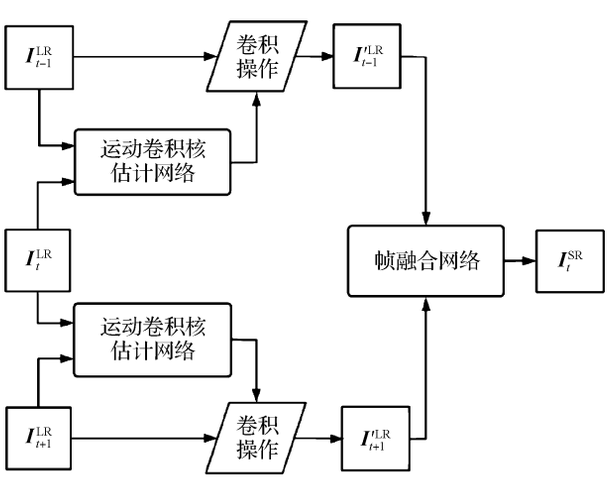

图17 RTVSR的网络结构（Bare等，2019）

Fig. 17 The network architecture of RTVSR (Bare et al., 2019)

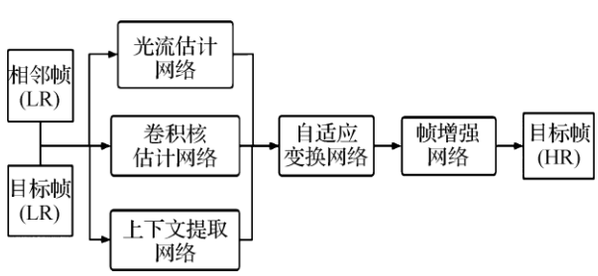

图18 MEMC-Net的网络结构（Bao等，2019b）

Fig. 18 The network architecture of MEMC-Net (Bao et al., 2019b)

在视频超分辨率领域，基于卷积核估计的配准方法如表6所示。由于卷积核大小限制运动估计的范围，这种方法的视频超分辨率重建效果相对较差。因此，单独使用卷积核估计来实现相邻帧对齐的方法相对较少。联合多种运动估计方法又会不可避免地带来昂贵的计算开销。

2.3 基于可变形卷积的配准方法

在运动估计与运动补偿过程中，光流法难以实现存在大尺度运动视频帧之间的配准。运动估计的不准确性会导致生成的高分辨率视频难以保证时间连续性。基于卷积核估计的运动范围受限于预定义

表 6　基于卷积核估计的配准方法

Table 6　The kernel-based alignment methods

方法	主要思想	放大倍率
RTVSR(Bare 等，2019)	运动卷积核估计	×2,×3,×4
MEMC-Net(Bao 等，2019b)	上下文提取网络	×4

内核的大小。额外的运动估计、补偿模块增加了网络学习的计算花销，限制了视频超分辨率方法在现实场景中的应用。因此，大量方法尝试使用可变形卷积实现相邻帧与目标帧的配准。

标准卷积对像素点进行规则采样，其固有的几何结构和固定的感受野难以适应视频帧内不同尺度、形状的目标。Dai 等人(2017)首次提出可变形卷积(deformable convolutional network)。可变形卷积通过对普通卷积中的采样点添加不同大小、方向的偏移量来改变感受野，使偏移后的感受野能覆盖图像内容，与实际形状更加匹配。

图 19 展示了可变形卷积的具体过程。输入特征图上的采样点通过额外的卷积核学习偏移量，且卷积核与当前卷积层具有相同的维度，输出的偏移量与输入的特征映射具有相同的空间分辨率。输出的通道维度为 $2N$，对应 N 个二维偏移量。网格中的采样位置通过学习的偏移量自由变形得到。

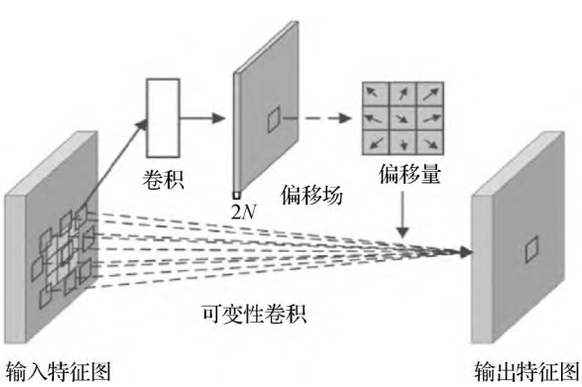

图 19　可变形卷积示意图(Dai 等，2017)

Fig. 19　Illustration of the sampling locations in standard and deformable convolutions(Dai et al. ，2017)

网格 $\boldsymbol{R}$ 定义了感受野的大小和空洞率，首先定义一个空洞率为 0 的 3 × 3 卷积核，感受野为

$$\boldsymbol{R}=\{(-1,-1),(-1,0),\cdots,(0,1),(1,1)\} \tag{11}$$

就传统卷积结构而言，对输出特征图 $\boldsymbol{y}$ 每个位置点 p_0 来说，都有

$$y(p_0)=\sum_{p_n\in R} w(p_n)\cdot x(p_0+p_n) \tag{12}$$

式中，符号 · 表示点乘，即向量的内积，p_n 是卷积输出每一个点相对感受野上的每一个点的偏移量，且 p_n 枚举了 $\boldsymbol{R}$ 中的位置。可变形卷积在式(12)的基础上，使网格 $\boldsymbol{R}$ 增加了偏移量 $\{\Delta p_n|n=1,\cdots,N\}$，其中，$N=|\boldsymbol{R}|$，则式(12)可改写为

$$y(p_0)=\sum_{p_n\in R} w(p_n)\cdot x(p_0+p_n+\Delta p_n) \tag{13}$$

现在，采样点的位置位于 $p_n+\Delta p_n$，使用双线性插值，利用周边采样点实现对偏移后位置的赋值。

可变形卷积对物体形变和建模能力较强，通过添加学习到的偏移量，可对感受野大小进行自适应改变以实现精确定位。同时，可变形卷积为轻量级模块，仅需少量的参数和计算量，即可轻易引入到现有网络中，已用于解决一系列高级视觉任务，例如目标检测(Dai 等，2017；Bertasius 等，2018)、语义分割(Dai 等，2017)和人体姿势估计(Sun 等，2018b)等。表 7 总结了基于可变形卷积配准的视频超分辨率方法。

表 7　基于可变形卷积的配准方法

Table 7　Deformable convolution-based alignment methods

方法	主要思想	放大倍率
TDAN(Tian 等，2020)	可变形卷积	×4
EDVR(Wang 等，2019b)	金字塔可变形卷积	×4
D3Dnet(Ying 等，2020)	3D 可变形卷积	×4
DNLN(Wang 等，2019a)	可变形卷积，non-local	×4
DKSAN(Xu 等，2020)	可变形卷积核	×16
VESR-Net(Chen 等，2020)	可变形卷积，Seperate non-local	×16
BasicVSR++(Chan 等，2021b)	光流指导可变形卷积	×4

时间可变形对齐网络(temporally deformable alignment network，TDAN)(Tian 等，2020)提出利用可变形卷积实现视频帧间的对齐。如图 20 所示，TDAN 主要由特征提取、可变形卷积对齐和对齐帧重构 3 部分组成。两个相邻帧经过多个残差卷积块构成的特征提取网络，获得相应的特征图。将特征图级联在一起，利用卷积网络学习可变形卷积核的各个采样位置偏移量。添加偏移量的卷积核可以实现相邻帧在特征级别上的对齐。然后，TDAN 将重

建的对齐帧特征与目标帧特征相结合，作为超分辨率重构网络的输入。可变形卷积取得了比利用SpyNet（Ranjan和Black，2017）估计光流进行运动补偿更好的对齐效果。之后，很多方法尝试改变可变形卷积的学习网络以及对齐过程，以此提升高分辨率视频帧的质量。

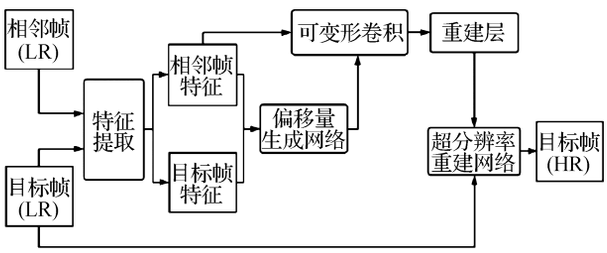

图20　TDAN的网络结构（Tian等，2020）

Fig. 20　The network architecture of TDAN（Tian et al.，2020）

3D卷积比2D卷积增加了时间维度的处理，可以同时建模外观与运动。TDAN将可变形卷积引入视频超分辨率的研究后，可以很自然地联想到将3D卷积与可变形卷积相结合，提出可变形3D卷积（Ying等，2020）的方法，进而针对视频超分辨率任务设计了D3Dnet（deformable 3d convolution video super-resolution）（Ying等，2020），其网络结构与TDAN相似，原有的可变形卷积被替换为3D可变形卷积。在空间维度执行卷积核变形的同时，引入时间先验，学习三维的偏移量，实现自适应的运动补偿。在计算成本合理增加的情况下实现了比TDAN更好的超分辨率性能。

Wang等人（2019b）运用三级金字塔和调制的可变形卷积（Zhu等，2019b）设计了PCD align module（pyramid，cascading and deformable alignment module）。首先，下采样原始视频序列。为了减少计算成本，使特征随空间大小降低而不增加通道数。其次，将对齐的低分辨率视频帧、偏移量进行上采样，与目标帧共同输入学习网络，以获得当前采样因子下的偏移量，并不断回溯，按照从粗略到精细的方式，不断提升可变形卷积对齐模块的精准度。

除了PCD align module，强化可变形卷积视频恢复网络（enhanced deformable video restoration，EDVR）利用预去模糊模块处理输入帧，提高原始视频特征质量，以避免低质量特征对后续的视频超分辨率网络产生负面效果，其网络结构如图21所示。一些方法提出将注意力机制引入网络中，设计了基于时间注意力与空间注意力的融合模块同时处理多幅对齐帧，然后重建融合特征图获得去模糊的、细节真实的高分辨率视频。值得一提的是，EDVR是NTIRE19视频恢复挑战赛（Nah等，2019a）中全部4个赛道的冠军模型。基于可变形卷积的视频超分辨率方法生成视频帧的质量在此次挑战赛中得到了证明。可变形卷积成为研究视频超分辨率网络的重要方法。

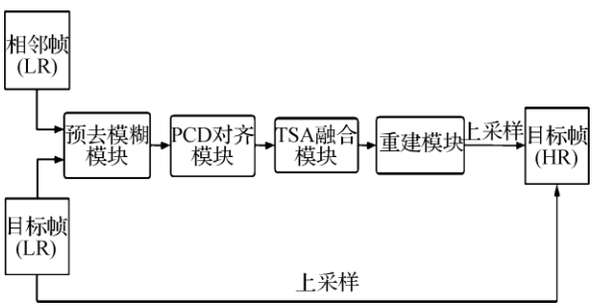

图21　EDVR的网络结构（Wang等，2019b）

Fig. 21　The network architecture of EDVR（Wang et al.，2019b）

另一些方法尝试对生成偏移量的学习网络进行改进，设计了基于层次特征融合块（hierarchical feature fusion block，HFFB）（Hui等，2021）的可变形非局部网络（deformable non-local network，DNLN）（Wang等，2019a）。HFFB利用空洞卷积（Yu和Koltun，2016）代替常规的2D卷积以获得更大的感受野，设计从1～8空洞率递增的卷积网络提取特征，将特征融合以获得感受野更加密集的卷积结果。再经过多个以此想法为基础的可变形卷积模块，实现从粗略到精细的偏移量学习过程。DNLN中还引入了non-local模块，该模块将在后文有更详细的介绍。如图22所示，在可变形卷积将相邻帧对齐处理后，DNLN利用对齐帧与参考帧的相关性增强视频帧缺失的细节信息。DNLN中将可变形卷积与non-local相结合的结构也是现有视频超分辨率网络的主流框架之一。

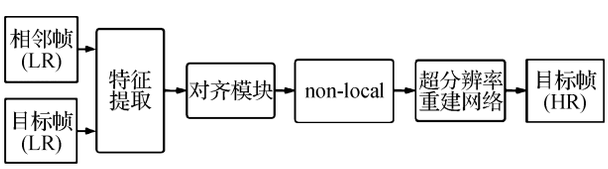

图22　DNLN的网络结构（Wang等，2019a）

Fig. 22　The network architecture of DNLN（Wang et al.，2019a）

与光流法相比，可变形卷积已经很大程度地改善了运动补偿效果。为了更为准确地提取运动信息，Gao 等人（2019）发现可变形卷积核可以通过学习卷积核的偏移量，对原卷积进行重新采样来适应有效感受野，将可变形卷积和可变形卷积核相结合可以提取更复杂的运动信息，对于分类任务和检测任务都十分高效。Xu 等人（2020）提出的可变形卷积核对齐模块（deformable kernel convolution alignment，DKC Align）可以提取整体轮廓，并改善局部边缘和纹理特征，同时对齐相邻帧。基于此模块的可变形卷积核空间注意网络（deformable kernel spatial attention network，DKSAN）最终获得了 AIM（artificial intelligence in medicine）2020 视频超分辨率挑战赛（Fuoli 等，2020）的第 2 名。

在 NTIRE 2021 挑战赛中，Chan 等人（2021a）发现可变形卷积对齐虽然具有优于光流对齐的性能，但是可变形卷积的训练过程难以收敛，偏移量的学习十分发散。根据形变对齐与光流对齐之间的强相关性，可利用 SpyNet（Ranjan 和 Black，2017）获取双向光流对齐相邻帧，并学习对齐后的相邻帧与目标帧间的偏移量，最终将光流与偏移量拼接在一起。光流引导的可变形卷积仅需学习较小的残差，减轻了常规形变对齐模块的负担。学习的调制掩码同时起到注意力机制的作用。BasicVSR++成功复原了图像的纹理细节，获得了 NTIRE 2021 视频超分辨率挑战赛的冠军。

与以前基于光流法的视频超分辨率不同，可变形卷积可以在特征级别自适应地对齐参考帧和相邻帧，而无需显式地进行运动估计和图像配准。对齐的低分辨率视频图像将具有较少的伪影，进而改善了重构的高分辨率视频质量。然而，可变形卷积本质是一种抽样改进，学习到的额外信息较少，且需要更多地关注偏移量的学习与优化。同时，如很多基于图像配准来进行视频超分辨率的技术一样，需要考虑如何利用对齐帧，以构建可以生成高质量、连续以及细节真实视频的超分重建网络。

2.4　基于图像配准方法的小结

无论是通过显式估计参考帧及其相邻帧之间的光流进行视频帧对齐，或是在特征级别隐式估计偏移量，利用可变形卷积对齐视频帧，都已经实现了优异的视频超分辨率效果。但是，额外的图像配准模块往往会带来昂贵的计算花销。在现实应用中，图像配准方法难以实现实时视频超分辨率的效果。且错误的配准方法会导致细节信息的缺失和错误特征的引入，对之后的融合重建过程产生负面作用，导致生成的视频中存在大量伪影。因此，可以考虑设计更精准的、轻量级的图像配准方法以改善视频超分辨率效率。

3　非图像配准的视频超分辨率方法

目前的图像配准模块无法适应实时性需求。相对地，许多研究不使用额外的图像配准模块，而是直接利用神经网络学习帧间的运动信息，融合重建时空信息的非图像配准的视频超分辨率方法。这些方法利用不同的网络结构隐式地利用视频帧间的关联性，从而更好地融合和重建出更清晰的纹理和丰富的细节。本节将对非图像配准视频超分辨率方法按照其基于的网络类型分别介绍。

3.1　基于3D卷积的非图像配准方法

3D 卷积（Tran 等，2015）是一种提取时空特征的有效手段，在视频分类和动作识别（Ji 等，2013）等领域有巨大的优势。在视频超分辨率问题上，3D-CNN 采用立体式的卷积核对整个视频序列进行卷积运算，可在时间维度上共享权重。相较 2D-CNN，3D-CNN 能更好地捕获视频的时空维度的特征信息。如图 23 和图 24 所示，2D-CNN 对单个视频帧在空间维度进行 2D 卷积操作以获得目标低分辨率帧 $\boldsymbol{I}_t^{\mathrm{LR}}$ 的特征 $\boldsymbol{F}_t$。3D-CNN 是对视频序列进行 3D 卷积操作，且卷积位置固定。图 24 中 3D-CNN 卷积的时间维度为 3，即对连续的 3 帧 $\boldsymbol{I}_{t-1}^{\mathrm{LR}}$，$\boldsymbol{I}_t^{\mathrm{LR}}$，$\boldsymbol{I}_{t+1}^{\mathrm{LR}}$ 进行时空维度卷积操作以获得目标帧的特征 $\boldsymbol{F}_t$。图中的直线表示对不同的帧的卷积操作，但是共享同一个卷积核。每组直线都对应卷积后提取的视频帧的特征图。卷积层中每一个特征图都会与时间维度对应的相邻帧相连，以此来捕捉帧间的运动信息。

表 8 总结了基于 3D 卷积的非图像配准方法。网络设计时，一般利用 3D 卷积进行单张视频帧空间信息的提取以及多张连续视频帧时空信息的融合。例如，3DSRNET（3D super-resolution network）（Kim 等，2019b）充分利用 3D 卷积来提取多帧的时空信息。3DSRNET 的网络结构如图 25 所示。图 25 中包含 6 个 3D 卷积层，每个卷积层中有 64 个 $3\times3\times3$ 的滤波器。随着网络的深入，3D 卷积后的特征图深

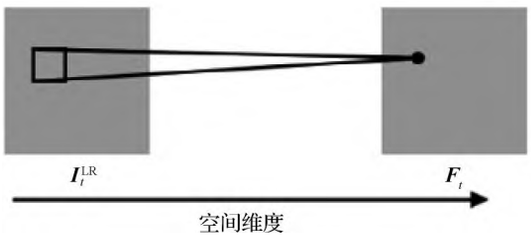

图23　2D卷积示意图(Tran等,2015)

Fig. 23　Illustration of the 2D convolution(Tran et al. ,2015)

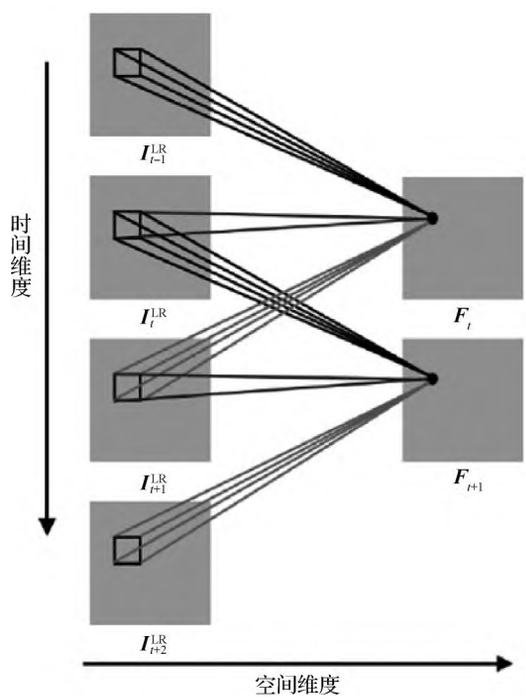

图24　3D卷积示意图(Tran等,2015)

Fig. 24　Illustration of the 3D convolution(Tran et al. ,2015)

表8　基于3D卷积的非图像配准方法

Table 8　3D convolution methods without alignment

方法	主要思想	放大倍率
DUF(Jo等,2018)	动态上采样滤波器	×2,×3,×4
3DSRNET(Kim等,2019b)	3D卷积	×2,×3,×4
FSTRN(Li等,2019b)	快速时空残差块	×4

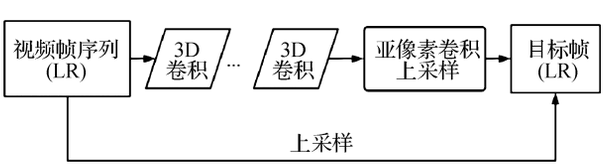

图25　3DSRNET的网络结构(Kim等,2019b)

Fig. 25　The network architecture of 3DSRNET (Kim et al. ,2019b)

度也会变浅。为了保持深度并保留时间信息,3DSRNET在卷积过程中,为连续视频帧的开头和结尾分别添加1帧。3DSRNET包含1个SF(scene change detection and frame replacement subnet)子网,对场景边界的确切位置进行分类,当检测到场景变化后,利用具有相同场景且最接近当前帧的图像进行替换,将替换后的序列发送到后续的视频超分辨率网络中。通过这种方法可以解决视频中场景变化的问题。

3D卷积层的过多利用会导致过量的计算负担与存储压力。FSTRN(fast spatio-temporal residual network)(Li等,2019b)利用低分辨率视频浅层特征提取网络(LR video shallow feature extraction net, LFENet),即3D卷积来获取每个视频帧独立的浅层空间特征,并在快速时空残差块(fast spatio-temporal residual block,FRB)中利用3D卷积提取时空信息,FRB模块用$1 \times k \times k$和$k \times 1 \times 1$的3D卷积层代替$k \times k \times k$卷积。将单个时空残差块上的3D卷积分解为两步,依次处理视频时空信息,可以很大程度上减轻计算压力。特征信息与FRBs的残差信息相融合之后,经过上采样超分辨率网络(LSRNet)生成高分辨率视频。

与常规利用3D卷积在时空维度提取连续帧的特征,然后融合重建的视频超分辨率方法不同,DUF(dynamic upsampling filters)(Jo等,2018)根据低分辨率帧中每个像素块的时空特征,生成局部的动态上采样滤波器。如图26所示,©表示连续帧的特征拼接,下同。⊗表示卷积操作。DUF首先对每个输入的连续帧利用共享的2D卷积层提取特征图,然后利用残差全连接的3D卷积网络构建的两个通道,分别获取残差信息与动态上采样滤波器。实验表明,该方法避免了显式运动估计补偿,对目标帧做动态卷积核上采样,是比传统的双线性插值法或双三次插值法更好地恢复图像纹理细节的方法。

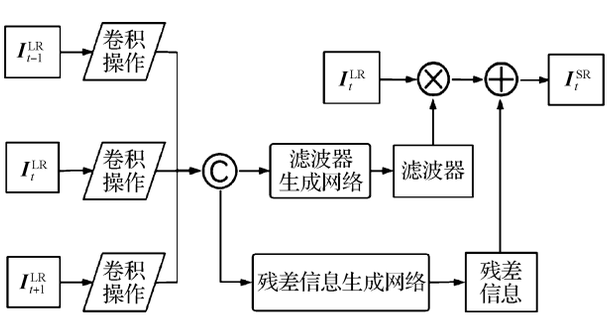

图26　DUF的网络结构(Jo等,2018)

Fig. 26　The network architecture of DUF(Jo et al. ,2018)

将3D卷积利用于视频超分辨率技术中，可以充分融合相邻帧与目标帧的时空信息，相对地，也会存在参数量大、运算速率慢的问题。如何提高3D卷积的计算效率并更好地提取与利用帧间运动信息是主要研究方向之一。

3.2　基于GAN的非图像配准方法

在前文中对GAN和基于GAN的光流配准方法进行了介绍，与其相同的是，无论是否使用光流，目前已有的对抗损失仅能提高单帧质量，难以保证视频连续性，并避免视频中的伪影。

Lucas等人(2019)将GAN的使用扩展到视频超分辨率问题，对损失函数进行了改进。随后，有方法提出在对抗损失的基础上，增加了内容损失和感知损失。对从生成图像和真实高分辨视频帧提取的特征，利用Charbonnier loss(Lai等，2017)代替最小化平方误差(MSE loss)计算差值。图27展示了VSRResFeatGAN(generative adversarial networks and perceptual losses for video super-resolution)的生成网络结构。在训练过程中，先训练由多个残差卷积块组成的生成网络，然后将判别器与生成器联合训练，可以得到较高质量的视频帧。

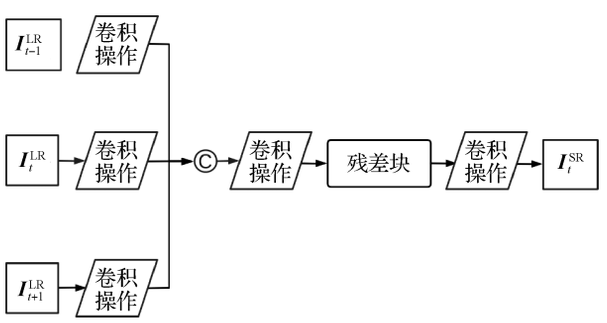

图27　VSRResFeatGAN的网络结构(Lucas等，2019)

Fig. 27　The network architecture of VSRResFeatGAN (Lucas et al. ，2019)

基于GAN的非图像配准方法如表9所示。可以看出，基于GAN进行视频超分辨率方法的研究非常少。因为鉴别器需要判断视频帧之间的连续性，因而难度较大。且在设计视频超分辨率网络时，生成器的结构过于简单，仅使用ResNet和DenseNet难以达到其他视频超分辨率方法的性能。

3.3　基于循环神经网络的非图像配准方法

RNN(Schuster和Paliwal，1997)和LSTM(Hochreiter和Schmidhuber，1997)在处理时序信息时展现出卓越效果。LSTM可以利用门结构控制单元状态，进而实现对特征信息的保护和控制。与RNN相比，这种结构在较长时序信息的处理上显示出明显优势。基本的门结构包括遗忘门、输入门和输出门。ConvLSTM(Shi等，2015；Toderici等，2016)提出将LSTM中权重相乘的方式替换成卷积操作，可实现图像序列的权重共享，进而更好地处理视频中的时空信息。由于视频时序信息的特点，在研究视频超分辨率时，需要考虑目标帧与前后帧之间的双向关系。Graves等人(2005)利用双向长短期记忆网络(bidirectional LSTM，B-LSTM)提取特征信息，之后融合重建时空特征来获得高分辨率视频。基于循环神经网络的非图像配准方法的主要思想和放大倍率如表10所示。

表9　基于GAN的非图像配准方法

Table 9　GAN methods without alignment

方法	主要思想	放大倍率
VSRResFeatGAN(Lucas等，2019)	感知损失	×2,×3,×4

表10　基于循环神经网络的非图像配准方法

Table 10　RNN methods without alignment

方法	主要思想	放大倍率
STCN(Guo和Chao，2017)	BMC-LSTM	×2,×3,×4,×5
BRCN(Huang等，2018)	循环神经网络，三维卷积	×2,×4
RISTN(Zhu等，2019a)	残差可逆块	×4
RLSP(Fuoli等，2019)	循环潜在信息	×4
RRN(Isobe等，2020c)	残差循环网络	×4
RSDN(Isobe等，2020a)	细节信息，结构信息	×4

时空卷积网络(spatio-temporal convolutional network，STCN)(Guo和Chao，2017)是最先利用ConvLSTM的视频超分辨率网络之一，网络结构如图28所示，其中的空间模块在每层卷积网络后利用PReLU(parametric rectified linear unit)(He等，2015)获取非线性响应。STCN利用时间模块处理连续视频特征中的时序信息。时间模块包含20个卷积层，每层由5个BMC-LSTM(bidirectional multi-scale convolutional LSTM)模块组成。如GoogleNet(Szegedy等，2015，2016)网络使用多个小型3 × 3卷积核替换大型卷积核，这样可以节省计算力。STCN合并输入帧

的特征与前一模块的输出信息，之后将特征放入3种不同深度的3×3卷积网络，然后合并不同通道的特征图，并利用1×1卷积网络压缩。双向网络可同时处理前后帧特征信息。将前向子网与后向子网的输出按照其时间通道合并在一起，并利用由一个卷积层组成的重建模块将时空信息直接映射至高分辨率视频帧。

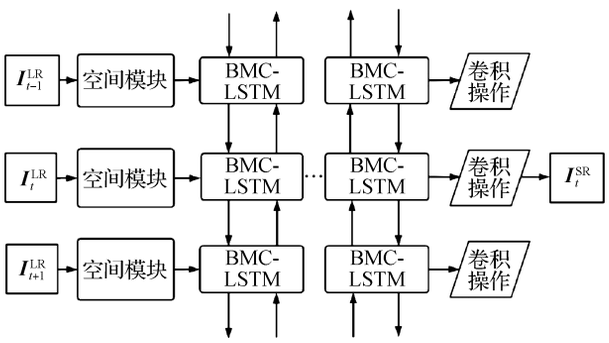

图28　STCN的网络结构（Guo和Chao，2017）

Fig. 28　The network architecture of STCN (Guo and Chao, 2017)

BRCN（bidirectional recurrent convolutional network）（Huang等，2018）将多个低分辨率连续帧分别输入到前馈子网与反馈子网，利用循环卷积神经网络来获取目标帧与前、后帧特征的时间关联。如图29所示，BRCN利用3D前馈网络连接输入层与隐藏层、连续隐藏层。而相邻帧之间的隐藏层则依靠循环卷积网络连接，当中的卷积核权值在所有时间内共享。这种方法可以减少网络的计算压力和存储负担。

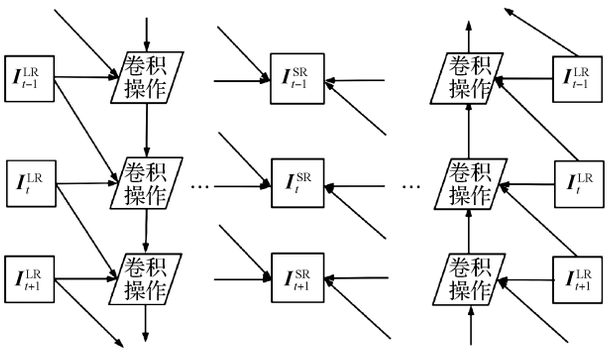

图29　BRCN的网络结构（Huang等，2018）

Fig. 29　The network architecture of BRCN (Huang et al., 2018)

RISTN（residual invertible spatial-temporal network）（Zhu等，2019a）与STCN的网络结构非常相似。RISTN在STCN基础结构上，更加充分地利用了残差网络和密集网络的特性。如图30所示，RISTN利用零填充层在RGB通道上构建连续帧的初始特征图。设计的残差可逆块将输入的特征图分为两部分。两部分的特征图并行经过多次卷积、归一化以及ReLU（rectified linear unit）函数后，合并作为时间网络模块的输入。RISTN在双向Conv-LSTM的基础上，结合残差网络和密集网络设计了RDC-LSTM（residual desnse convolution LSTM）。时间网络部分由多个RDC-LSTM组成，可以更好地提取前后帧与目标帧间的时序关系，并且研究人员提出了一种稀疏特征融合的方法选择特定的特征图，以此保证重建视频的质量。

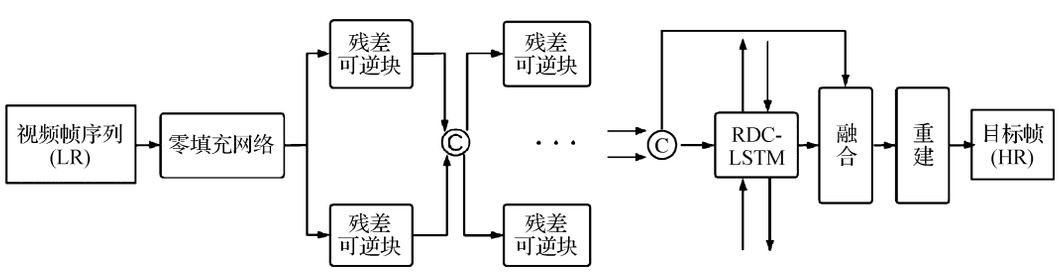

图30　RISTN的网络结构（Zhu等，2019a）

Fig. 30　The network architecture of RISTN (Zhu et al., 2019a)

RLSP（recurrent latent space propagation）（Fuoli等，2019）利用亚像素卷积的可逆性，同时将连续3帧下采样过的上一高分辨率输出帧和前一隐藏状态的输出作为当前循环隐藏状态的输入，以有效地隐式利用时间信息。如图31所示，RLSP的网络结构轻便，不需要额外的运动估计与运动补偿模块，参数量较少。与利用光流运动估计的FRVSR（Sajjadi等，2018）和生成动态滤波器的DUF（Jo等，2018）相比，RLSP分别实现了约10倍和70倍的加速，同时在Vid4数据集上保持了相似的精度。

Isobe等人（2020c）针对RLSP进行了两点改进，实现重建视频在质量上的提升。1）如图32所示，将

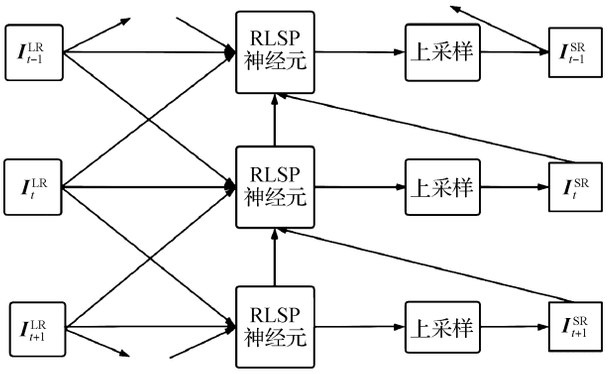

图 31　RLSP 的网络结构（Fuoli 等，2019）

Fig. 31　The network architecture of RLSP (Fuoli et al., 2019)

两个连续帧而不是 3 个连续帧送入每个隐藏层，以减少错误累计。2）在隐藏层中添加了残差学习，以减少梯度消失的影响。MAI（mobile AI）2021 实时视频超分辨率挑战赛中，Liu 等人（2021）在 RNN 基础上利用 NAS（neural architecture search）（Kim 等，2019a）进行网络优化，获得了挑战赛的第 2 名。

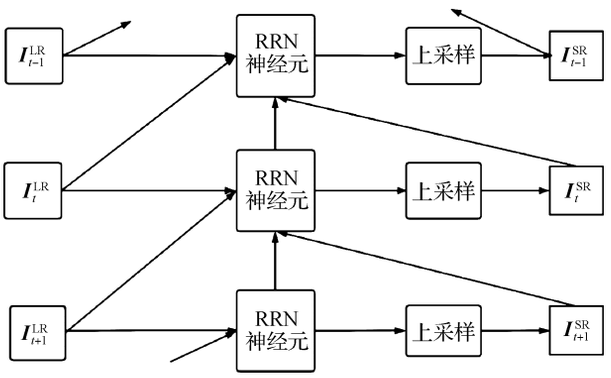

图 32　RRN 的网络结构（Isobe 等，2020c）

Fig. 32　The network architecture of RRN (Isobe et al., 2020c)

Isobe 等人（2020a）提出的 RSDN（recurrent structure-detail network）将视频帧分为细节部分的高频信息和结构部分的低频信息，这两部分被分别馈送到由多个结构—细节块（structure-detail block）组成的循环单元中，并在网络中设计了隐藏状态适配模块（hidden state adaptation），允许当前帧有选择地使用来自隐藏状态的信息，以实现更好的性能和更少的错误累积风险。利用循环网络的特性对高低频信息分别处理，可以增强对外观变化和细节模糊等问题的鲁棒性。

Yi 等人（2021）设计的 OVSR（omniscient video super-resolution）在循环网络的思路上进行思考，进一步设计了一种全能（omniscient）框架。OVSR 与 RLSP 等方法相同，将包括目标帧的连续 3 帧作为输入，并利用了前一帧的超分辨率结果，但它同时利用了当前帧与后一帧的超分辨率结果。在对两重建子网进行设计时，沿用了 PFNL 的融合框架，利用 3 个分支对视频序列的空间相关性、时序相关性进行充分探索。OVSR 在性能上大幅超越包括 EDVR 在内的很多图像配准方法，展现了非图像配准方法可以达到推理速度和视觉质量的平衡。

循环神经网络可以充分利用帧间互补信息，对视频的时序信息进行建模，以实现有效的视频超分辨率。目前基于循环神经网络的模型参数量都较为适中，计算负担普遍小于基于运动估计、运动补偿的方法和基于 3D 卷积的方法。OVSR 证明了循环神经网络在视频超分辨率领域的潜能，因此可以更多地尝试调整隐藏层网络，引入更合适的注意力融合机制，实现更出色的视频超分辨率效果。

3.4　基于 non-local 的非图像配准方法

Wang 等人（2018b）提出的 non-local Net 在很大程度上促进了计算机视觉领域的发展。其中的核心非局部模块（non-local block）可以很容易地融入到现有的深度神经网络中，有效处理时空相关性信息。因此，在解决时空维度的问题，例如视频分类、目标检测与分割、姿态估计等任务时，non-local 都能带来显著的效果。

非局部模块主要针对长程相关性（long-range dependence，LRD）问题进行设计。当使用常规卷积操作时，感受野都会受到卷积核大小的限制，仅从局部区域考虑问题。但在非局部模块中不同，其可以从全局获取各特征的相关性。具体为

$$\boldsymbol{y}_i = \frac{1}{C(\boldsymbol{x})}\sum^{\forall j} f(\boldsymbol{x}_i, \boldsymbol{x}_j) g(\boldsymbol{x}_j) \tag{14}$$

式中，$\boldsymbol{x}$ 表示输入数据（图像、视频和文本等），$\boldsymbol{y}$ 表示大小与 $\boldsymbol{x}$ 相同的输出。i 是输出位置的索引，j 是所有可能位置的索引。函数 $f(\cdot)$ 用来计算一个表示输入之间某种关系的标量，例如高斯函数、点乘函数等。而 $g(\cdot)$ 表示处理输入的映射函数，$C(x)$ 用于标准化。

在视频超分辨率研究中，需要考虑视频帧之间的运动关系。non-local 可以计算目标帧特征与所有相邻帧特征间的相关性。在加权平均和计算所有位置之间响应的过程中，可以充分利用时空信息对连

续帧进行处理。因此，non-local常作为运动估计与运动补偿方法的替代，近两年提出的所有基于non-local的视频超分辨率算法都获得显著成效，经典方法包括PFNL和MuCAN，其主要思想和放大倍率如表11所示。

表11　基于non-local的非图像配准方法

Table 11　The non-local methods without alignment

方法	主要思想	放大倍率
PFNL(Yi等，2019)	non-local	×4
MuCAN(Li等，2020)	帧间相似性，帧内相似性	×4

Yi等人(2019)提出的渐进融合非局部网络(progressive fusion non-local，PFNL)是最先使用非局部残差块(non-local residual block，NLRB)来处理视频的时空相关性的方法之一，如图33所示，该网络将经过NLRB处理的特征图放入渐进融合模块(progress fusion resblock，PFRB)进行特征融合。首先，PFRB分别提取连续帧的特征图，以获取各自独立的空间信息，将空间信息合并后，利用卷积操作获得混合的时空特征。然后，将生成的特征图与各自独立的空间信息放在一起进行卷积，并将最后的结果依靠残差网络的特性添加到输入帧之中。实验证明了NLRB可以充分利用相邻特征与目标帧像素级别的相关性，并增强所需的缺失细节。相比于运动估计与运动补偿的方法，PFNL不需要额外的光流网络或偏移量学习网络，可以减少网络中的参数量，并且不需要额外的损失函数。

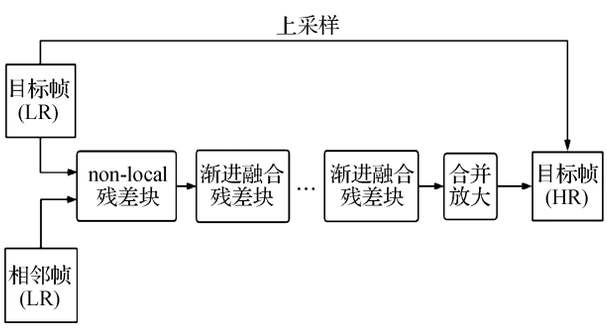

图33　PFNL的网络结构(Yi等，2019)

Fig. 33　The network architecture of PFNL(Yi et al.，2019)

non-local从全局捕获各特征的相关性，而视频帧间和帧内本存在很多相似信息。基于此，MuCAN(multi-correspondence aggregation network)(Li等，2020)设计了两个有效模块，以充分利用内容相似性超分辨率目标帧。时间多相关性集成模块(temporal multi-correspondence aggregation module，TM-CAM)首先将目标帧与相邻帧以较低分辨率编码，然后计算编码特征的相似性，在局部选择最相似的多个特征块，并利用像素自适应聚合方法，即在不同位置生成权重，以启用不同的聚合模式将它们融合成一个像素。融合特征保持了亚像素级别的细节，上采样后可以从粗到细地聚合高分辨率特征。交叉尺度非局部相关性集成模块(cross scale nonlocal correspondence aggregation module，CN-CAM)构建特征金字塔，探索多个空间尺度视频帧的相似特征，并利用自注意力机制来选取特征信息。该模块使网络能够揭示更多细节，进一步增强重建视频的质量。在训练时，MuCAN使用边缘感知损失以避免产生锯齿状边缘，并能确保生成清晰的纹理。

传统的non-local block计算成本非常高，而且需要存储包含庞大参数的注意力矩阵。一些基于图像配准的方法对non-local进行改进，并提出新颖的模型，例如前文介绍的DNLN和VESR-Net。Chen等人(2020)设计了一种称为分离非局部模块(seperate non-local)的结构。Seperate non-local包含3种类型的注意力模块，分别探索不同空间上下文、不同通道和不同时间步长之间的相似性，跨所有视频帧融合像素信息，并且引入了通道注意力，以此实现高效的视频帧重建。VESR-Net可以在网络运算负担相对较小的情况下达到很好的性能。在2020年优酷视频挑战赛中，VESR-Net排名第1。

基于non-local模块的视频超分辨率技术取得了显著效果，与网络更加复杂的基于运动估计、运动补偿的视频超分辨率方法相比，non-local在参数量和生成视频质量等方面均展现出优势，但非局部模块同时处理多个视频帧的计算负荷昂贵，且在视频超分辨率任务中，当前方法使用的尺寸较大，具有较高分辨率的视频帧。如果增加输入的连续视频帧数量和特征维度，无疑是在指数级别提高计算量需求。因此，如何对non-local模块更好地分割和利用是视频超分辨率及相关领域的重大难题。

3.5　非图像配准方法的小结

不使用图像配准的视频超分辨方法在时间效率上显示出较大优势，神经网络可以相互耦合地学习序列图像的时空信息，但由于并未进行视频帧的配准，一些非目标帧中的低质量特征也会参与融合，可

能重建出模糊的纹理和错误的细节。有复杂纹理区域的视频帧会导致网络学习的难度过大。因此，可以尝试引入更高效的注意力机制，设计更精妙的特征融合模块，增加对非目标帧特征的鲁棒性，以改善多帧融合的效果。

4　实验分析

自 2015 年 Liao 等人（2015）首次提取深度特征应用到视频超分辨率领域以来，基于深度学习的视频超分辨率算法层出不穷，包括 EDVR（Wang 等，2019b）、VESPCN（Caballero 等，2017）、TDAN（Tian 等，2020）和 VESR-Net（Chen 等，2020）等。本文对以标准卷积神经网络、三维卷积神经网络、循环卷积神经网络、生成式对抗网络以及其他网络模型为基础构建的视频超分辨率方法进行了介绍。本节在多个标准测试数据集下，对部分网络模型在 4 倍放大因子下的 PSNR（/dB）/SSIM 值进行比较。使用两种 4 倍下采样方法 BI（bicubic）和 BD（blur downsampling）获取低分辨率视频序列，进行模型的测试和比较。实验数据都源于对开源模型的测试和文献中公布的结果。

表 12—表 14 展示了多种经典视频超分辨率网络在 REDS、Vimeo-90K-T、Vid4 和 UDM10 等 4 个常用数据集上的 PSNR/SSIM 值。表 15 对一些尚未开源且采用测试集使用范围较小的视频超分辨率方法进行了总结。表 16 展示了各经典视频超分辨率网络的内存开销、计算量和运行时间。运行时间为各网络对 180 × 320 像素的低分辨率帧进行 4 倍超分辨率至 720 ×1 280 像素的高分辨率帧花费的时间。由于部分模型尚未开源或主要针对 16 倍超分辨率进行设计而难以公平比较，因此缺少部分数据。利用这些测量数据，本文对主流方法进行充分比较和分析。

表 12　基于光流的网络模型在 REDS、Vimeo-90K-T、Vid4 和 UDM10 测试集上×4 倍率重构结果

Table 12　Results of the optical flow-based methods on REDS，Vimeo-90K-T，Vid4 and UDM10 with scale factor ×4

方法	PSNR(/dB)/SSIM					
	REDS（BI）	Vimeo-90K-T（BI）	Vid4（BI）	UDM10（BD）	Vimeo-90K-T（BD）	Vid4（BD）
Bicubic（Dong 等，2016）	26.14/0.729 2	31.32/0.868 4	23.78/0.634 7	28.47/0.825 3	31.30/0.868 7	21.80/0.524 6
VSRnet（Kappeler 等，2016）	–	32.43/0.889 3	24.37/0.679 3	–	–	–
VESPCN（Caballero 等，2017）	–	33.55/0.907 1	25.35/0.730 9	–	–	–
MMCNN（Wang 等，2019c）	–	–	26.28/0.784 4	–	–	–
DRVSR（Tao 等，2017）	–	–	25.88/0.775 2	–	–	–
MultiBoot（Kalarot 和 Porikli，2019）	31.00/0.882 2	–	–	–	–	–
SOFVSR（Wang 等，2018a）	–	34.89/0.923 4	26.01/0.771 0	–	–	–
MEMC-Net（Bao 等，2019b）	–	33.47/0.947 0	–	–	–	–
TecoGAN（Chu 等，2020）	–	–	–	–	–	25.57/0.757 2
RRCN（Li 等，2019a）	–	–	25.86/0.759 1	–	–	–
TOFlow（Xue 等，2019）	27.98/0.799 0	33.08/0.905 4	25.89/0.765 1	36.26/0.943 8	34.62/0.921 2	–
FRVSR（Sajjadi 等，2018）	–	–	–	37.09/0.952 2	35.64/0.931 9	26.69/0.810 3
RBPN（Haris 等，2019）	30.09/0.859 0	37.07/0.943 5	27.12/0.818 0	38.66/0.959 6	37.20/0.945 8	–
BasicVSR（Chan 等，2021a）	31.42/0.890 9	37.18/0.945 0	27.24/0.825 1	39.96/0.969 4	37.53/0.949 8	27.96/0.855 3
IconVSR（Chan 等，2021a）	31.67/0.894 8	37.47/0.947 6	27.39/0.827 9	40.03/0.969 4	37.84/0.952 4	28.04/0.857 0
VSR Transformer（Cao 等，2021）	31.19/0.881 5	37.71/0.949 4	27.36/0.825 8	–	–	–
BasicVSR++（Chan 等，2021b）	32.39/0.906 9	37.79/0.950 0	27.79/0.840 0	40.72/0.972 2	38.21/0.955 0	29.04/0.875 3

注：“–”表示该方法代码未开源或未在相对应的数据集训练。

表13　基于可变形卷积的网络模型在REDS、Vimeo-90K-T、Vid4和UDM10测试集上×4倍率重构结果

Table 13　Results of the deformable convolution methods on REDS, Vimeo-90K-T, Vid4 and UDM10 with scale factor ×4

方法	PSNR(/dB)/SSIM					
	REDS(BI)	Vimeo-90K-T(BI)	Vid4(BI)	UDM10(BD)	Vimeo-90K-T(BD)	Vid4(BD)
Bicubic(Dong等,2016b)	26.14/0.729 2	31.32/0.868 4	23.78/0.634 7	28.47/0.825 3	31.30/0.868 7	21.80/0.524 6
TDAN(Tian等,2020)	–	35.34/0.930 8	26.24/0.779 0	–	–	–
D3Dnet(Ying等,2020)	–	35.65/0.933 5	26.52/0.799 0	–	–	–
EDVR-M(Wang等,2019b)	30.53/0.869 9	37.09/0.944 6	27.10/0.818 6	39.40/0.966 3	37.33/0.948 4	27.45/0.840 6
DNLN(Wang等,2019a)	–	37.38/0.947 3	27.29/0.824 7	–	–	–
EDVR(Wang等,2019b)	31.09/0.880 0	37.61/0.948 9	27.35/0.826 4	39.89/0.968 6	37.81/0.952 3	27.85/0.850 3
BasicVSR++(Chan等,2021b)	32.39/0.906 9	37.79/0.950 0	27.79/0.840 0	40.72/0.972 2	38.21/0.955 0	29.04/0.875 3

注:“–”表示该方法代码未开源或未在相对应的数据集训练。

表14　非图像配准网络模型在REDS、Vimeo-90K-T、Vid4和UDM10测试集上×4倍率重构结果

Table 14　Results of the super-resolution methods without alignment on REDS, Vimeo-90K-T, Vid4 and UDM10 with scale factor ×4

方法	PSNR(/dB)/SSIM					
	REDS(BI)	Vimeo-90K-T(BI)	Vid4(BI)	UDM10(BD)	Vimeo-90K-T(BD)	Vid4(BD)
Bicubic(Dong等,2016b)	26.14/0.729 2	31.32/0.868 4	23.78/0.634 7	28.47/0.825 3	31.30/0.868 7	21.80/0.524 6
3DSRnet(Kim等,2019b)	–	–	25.71/0.758 8	–	–	–
VSRResFeatGAN(Lucas等,2019)	–	–	25.51/0.753 0	–	–	–
FRVSR(Sajjadi等,2018)	–	–	–	37.09/0.952 2	35.64/0.931 9	26.69/0.810 3
DUF(Jo等,2018)	28.63/0.825 1	–	–	38.48/0.960 5	36.87/0.944 7	27.38/0.832 9
PFNL(Yi等,2019)	29.63/0.850 2	36.14/0.936 3	26.73/0.802 9	38.74/0.962 7	–	27.16/0.835 5
MuCAN(Li等,2020)	30.88/0.875 0	37.32/0.946 5	–	–	–	–
RLSP(Fuoli等,2019)	–	–	–	38.48/0.960 6	36.49/0.940 3	27.48/0.838 8
RISTN(Zhu等,2019a)	–	–	26.13/0.79 2	–	–	–
RSDN(Isobe等,2020a)	–	–	–	39.35/0.965 3	37.23/0.947 1	27.92/0.850 5
RRN(Isobe等,2020c)	–	–	–	38.96/0.964 4	–	27.69/0.848 8
OVSR-S(Yi等,2021)	–	–	–	39.37/0.967 3	–	27.99/0.859 9
OVSR-L(Yi等,2021)	–	–	–	40.14/0.971 3	37.63/0.950 3	28.41/0.872 4

注:“–”表示该方法代码未开源或未在相对应的数据集训练。

表15　网络模型在非常规测试集上重构结果

Table 15　Comparision results of the video super-resolution methods on rare datasets

方法	测试集	下采样方式	放大倍率	PSNR/dB	SSIM	方法类别
TecoGAN(Chu等,2020)	ToS	Blur Downsampling	×4	32.75	–	GAN,光流配准
STARnet(Haris等,2020)	Middlebury	Bicubic	×4	27.16	0.827 0	可变形卷积配准
RVSR(Liu等,2017)	UVGD	Bicubic	×4	39.71	–	常规卷积,光流配准
FSTRN(Li等,2019b)	TDTFF	Blur Downsampling	×4	29.95	0.870 0	3D卷积,非图像配准
BRCN(Huang等,2018)	TDTFF	Blur Downsampling	×4	28.20	0.773 9	循环卷积,非图像配准
STCN(Guo和Chao,2017)	Hollywood2	Blur Downsampling	×4	34.58	0.925 9	循环卷积,非图像配准
VESR-Net(Chen等,2020)	YouKu VESR-T	Blur Downsampling	×16	35.97	–	可变形卷积配准
DKSAN(Xu等,2020)	IntVID	Blur Downsampling	×16	31.43	–	可变形卷积配准

注:“–”表示该方法代码未开源或未在相对应的数据集训练。

表 16　各个网络模型的训练帧数、参数量、运行时间、训练集、实验平台及损失函数

Table 16　Frames, parameters, running time, dataset, experimental platform and loss function of each method

方法	处理帧数	参数量/M	运行时间/ms	训练集	实验平台	损失函数
Bicubic(Dong等,2016b)	1	N/A	N/A	N/A	N/A	N/A
VESPCN(Caballero等,2017)	3	0.88	–	CDVL	–	MSE loss+MC loss
3DSRnet(Kim等,2019b)	5	–	–	largeSet	Titan X GPU	MSE loss
VSRResFeatGAN(Lucas等,2019)	5	–	–	Myanmar	Titan X GPU	Charbonnier loss + adversarial loss
MMCNN(Wang等,2019c)	7	10.5	–	HD documentaries	1080Ti GPU	MSE loss+MC loss
DRVSR(Tao等,2017)	3	6.6	–	SPMC	Titan X GPU	MSE loss+MC loss
MultiBoot(Kalarot和Porikli,2019)	7	60	200	REDS	Titan X GPU	Huber loss
SOFVSR(Wang等,2018a)	3	1.64	250	CDVL	970 GPU	MSE loss+MC loss
RTVSR(Bare等,2019)	3	15	–	monicinc	–	MSE loss
MEMC-Net(Bao等,2019)	3	67.2	640	Vimeo-90K	Titan X GPU	Charbonnier loss
TOFlow(Xue等,2019)	7	1.41	1 610	Vimeo-90K	Titan X GPU	L1 loss
FRVSR(Sajjadi等,2018)	recurrent(2)	5.1	137	vimeo.com	–	MSE loss+MC loss
DUF(Jo等,2018)	7	5.8	974	Internet	1080Ti GPU	Huber loss
TecoGAN(Chu等,2020)	7	3	–	SPMC	Titan X GPU	Ping-Pong loss
RBPN(Haris等,2019)	7	12.2	1 507	Vimeo-90K	Titan X GPU	L1 loss
TDAN(Tian等,2020)	5	1.97	–	Vimeo-90K	–	L1 loss
D3Dnet(Ying等,2020)	7	2.58	–	Vimeo-90K	2080Ti GPU	MSE loss
EDVR-M(Wang等,2019b)	7	3.3	118	Vimeo-90K+REDS	Titan X GPU	Charbonnier loss
EDVR(Wang等,2019b)	7	20.6	378	Vimeo-90K+REDS	Titan X GPU	Charbonnier loss
PFNL(Yi等,2019)	7	3	295	MM522	Titan X GPU	Charbonnier loss
MuCAN(Li等,2020)	7	–	–	Vimeo90K	Titan X GPU	Edge-aware loss
TGA(Isobe等,2020b)	7	5.8	375	Vimeo90K	Tesla V100 GPU	L1 loss
RLSP(Fuoli等,2019)	recurrent(3)	4.2	49	Vimeo90K	Titan X GPU	MSE loss
RSDN(Isobe等,2020a)	recurrent(2)	6.2	94	Vimeo-90K	Tesla V100 GPU	Charbonnier loss
RRN(Isobe等,2020c)	recurrent(2)	3.4	45	Vimeo-90K	Titan X GPU	L1 loss
VESR-Net(Chen等,2020)	7	15.96	–	YouKu VESR	1080Ti GPU	–
DKSAN(Xu等,2020)	7	29.5	–	Vid3oC	Titan X GPU	Charbonnier loss
VSR Transformer(Cao等,2021)	7	43.8	–	Vimeo-90K+REDS	Titan X GPU	Charbonnier loss
OVSR-S(Yi等,2021)	recurrent(3)	1.9	25.4	MM522	1080Ti GPU	Charbonnier loss
OVSR-L(Yi等,2021)	recurrent(3)	7.1	81.2	MM522	1080Ti GPU	Charbonnier loss
BasicVSR(Chan等,2021a)	recurrent(3)	6.3	63	Vimeo-90K+REDS	Tesla V100 GPU	Charbonnier loss
IconVSR(Chan等,2021a)	recurrent(3)	8.7	70	Vimeo-90K+REDS	Tesla V100 GPU	Charbonnier loss
BasicVSR++(Chan等,2021b)	recurrent(5)	7.3	77	Vimeo-90K+REDS	Tesla V100 GPU	Charbonnier loss

注:“–”表示该方法代码未开源或未公布相应数据。

4.1 基于卷积核估计的配准方法实验分析

由于基于卷积核估计的经典图像配准方法较少，仅包含MEMC-Net（Kalarot和Porikli，2019）和RTVSR（Bare等，2019）两种，且MEMC-Net属于基于光流配准的方法，因此未单独列表对比，在表12中列举了MEMC-Net的实验结果。RTVSR在Bicubic下采样的Vid4上PSNR/SSIM测试结果为26.36 dB/0.79。MEMC-Net的参数量高达67.2 M，重建720×1像素280视频帧更是需要640 ms。在Vimeo-90K上，MEMC-Net的测试结果仅为33.47 dB /0.947 0，效果劣于参数量远少于它的EDVR（Wang等，2019b）和BasicVSR（Chan等，2021a）等网络。基于卷积核估计的视频超分辨率方法效果较差，且之后少有学者尝试使用基于卷积核估计的图像配准方法进行视频超分辨率的研究。

4.2 基于光流的配准方法实验分析

基于光流的视频超分辨率方法数量最多，其中基于常规卷积的光流配准方法发展时间最久。最初的研究未使用统一测试集，但都集中于常规下采样低分辨率视频的实验，未考虑模糊核等在降质过程的作用。TOFlow（Xue等，2019）在REDS和Vid4上取得了高效性能，表明Vimeo90K数据集对网络训练具有积极作用。Vimeo90K也成为其他视频超分辨率网络训练、测试时使用的主流数据集。基于常规卷积的光流配准方法已逐渐淘汰，在网络结构设计上有愈来愈多的创新。基于GAN的方法TecoGAN（Chu等，2020）在Vid4上测试时，PSNR/SSIM指标为25.57 dB/0.757 2，效果不尽如人意，但在设置的tOF和tLP评价指标上的质量分数与EDVR不相上下，这两项指标更能反映帧间的连贯性。这也证明基于GAN的方法在真实视频超分辨率和盲视频超分辨率的研究中具有巨大潜力。

基于循环神经网络的光流配准方法发展迅速，最初的方法RRCN（Li等，2019a）的重建效果不佳，其仍是在图像层面配准相邻帧，将配准后的相邻帧与目标帧作为循环网络的输入。RBPN将光流信息、目标帧与相邻帧作为输入，在特征级别处理多种信息，对循环网络的输出利用反投影网络重建高分辨率目标帧，获得了一定的成功。但RBPN的参数量为12.2 M，其计算复杂度带来的运算负担更是难以承受，重建720×1 280像素视频帧需要1.507 ms。

上述方法未摆脱滑动窗口法的多帧重建单帧的思想，难以发挥循环神经网络的优势。而最近Chan团队的BasicVSR（Chan等，2021a）取得了视频超分辨率领域的重大突破。BasicVSR的参数量仅为6.3 M，在REDS、Vimeo-90K-T、Vid4和UDM10测试集上都刷新了PSNR/SSIM的质量分数。基于BasicVSR改进的IconVSR（Chan等，2021a）和BasicVSR++（Chan等，2021b）都进一步展现了有效性。这3种方法都是利用光流配准前一隐藏层的输出特征，运用循环神经网络处理低分辨率视频序列，一次性输出所有高分辨率视频帧。参数量少，视频超分辨率效果突出，而且这种巧妙的网络设计还有广阔的提升空间。循环神经网络与光流法相结合的方式可能如BasicVSR的名字一样成为未来很多研究的基准。

基于Transformer的视频超分辨率网络仅有VSR Transformer（Cao等，2021），方法太少，无法判断Transformer对该任务是否完全适配。虽然VSR Transformer在性能上达到很多顶尖方法水准，多项指标超越EDVR，在Vimeo90K上Bicubic下采样的测试结果PSNR/SSIM值为37.71 dB/0.949 4，略低于BasicVSR++，但随之带来的是巨大的计算压力和参数量，其参数量达到43.8 M，与VSR Transformer在性能上相差不多的基于循环神经网络的方法在参数量上有显著差距。但Transformer此前已经被证明能很好地处理类似视频的序列信息（音频、文字），相信未来能减少视频超分辨率Transformer的计算复杂度，或在如此参数规模上展现Transformer的性能优势。

4.3 基于可变形卷积的配准方法实验分析

NTIRE 2019视频恢复4个赛道冠军EDVR和NTIRE 2021视频恢复三冠一亚的方案BasicVSR++都是基于可变形卷积的图像配准方法，这证明了可变形卷积在视频超分辨率领域的有效性。2018年，TDAN最先使用可变形卷积，在BI下采样的Vid4上PSNR/SSIM值达到26.24 dB/0.779 0的效果，当时并未引起注意。但EDVR方法提出以后，可变形卷积在视频超分辨率领域的应用开始流行起来。EDVR的参数量为20.6 M，与之前的方法相比很庞大，且复现时往往效果不好，可变形卷积的训练一直难以收敛，需要多次断点训练和调参，但这些没能阻止可变形卷积成为各大挑战赛的常用方法。进行16倍视频超分任务时，可变形卷积对齐也发挥了突出作用。VESR-Net获得了YouKu 2020视频超分挑战赛

冠军，在YouKu VESR-T数据集上PSNR达到35. 97 dB。DKSAN获得了AIM 2020视频超分辨率挑战赛的亚军，在IntVID数据集上PSNR达到31. 43 dB。运用non-local模块和可变形卷积对齐模块为基础的超分辨率重构模型越来越多，重建视频的质量也能得到保证。但值得注意的是，参数量和计算量过于庞大阻止了此类模型在现实场景中的应用。

基于可变形卷积的BasicVSR++在7. 3 M的参数量下实现了出色的重建效果，在多个数据集下都展现了SOTA（state-of-the-art model）的性能。BasicVSR++在UDM10数据集中的PSNR为40. 72 dB，在Vid4上实现26. 24 dB的效果。可变形卷积方法的高效性和潜力得到了展示。如前文所述，与其他方法相比，BasicVSR++的最大优势为光流指导的可变形卷积。将可变形卷积与光流相结合的方式减轻了偏移量的学习负担，偏移量可以视为光流的残差信息，两者相辅相成。

可变形卷积的偏移量学习过程可以进一步改进，例如利用多维度的空洞卷积扩大感受野，或增加注意力机制提升偏移量的准确性。未来，可变形卷积在视频超分辨率中还有巨大潜能。

4. 4　非图像配准方法实验分析

MEMC-Net（Bao 等，2019）、EDVR（Wang 等，2019b）和VESR-Net（Chen等，2020）等视频超分辨率网络都包含额外的图像配准模块，计算压力过重。与上述图像配准方法相比，非图像配准方法最大的优势是参数量少、运行时间短，能够满足实时性的需求。

本文将非图像配准视频超分辨率方法按照网络结构分为4种。最能突出其及时性优势的为RLSP（Fuoli 等，2019）、RRN（Isobe 等，2020c）和RSDN（Isobe等，2020a）等基于循环神经网络的非图像配准方法。在超分辨率重构速度上，RLSP和RRN网络重建720 × 1 280像素视频帧的速度分别为49 ms和45 ms，OVSR轻量级版本的重建速度为25. 4 ms，在模糊核下采样的UDM10和Vid4数据集上达到了非图像配准视频超分辨率方法的最佳性能，在Vid4上的PSNR/SSIM为27. 99 dB/0. 859 9，超过了EDVR和BasicVSR等性能优越的基于图像配准的视频超分辨率方法。OVSR的重量级版本在多个数据集上达到了与BasicVSR++相近的性能。这证明了非图像配准视频超分辨率方法的潜力。

基于3D卷积的非图像配准方法较少，当中性能最佳的方法为DUF（Jo等，2018），其在UDM10数据集上的PSNR/SSIM为38. 48 dB/0. 960 5，与参数量较小的RLSP几乎一样，但速度却比RLSP慢了70倍，需要974 ms。DUF需要估计每个位置的动态卷积核，并且使用了大量3D卷积，计算量极其庞大，在处理大尺寸视频帧时对显存需求很高。基于3D卷积的非图像配准方法失去了其速度的优势，3D卷积直接融合多帧特征也难以达到基于图像配准方法的性能。

基于GAN的非图像配准方法VSRRes- FeatGAN（Lucas 等，2019）的性能较差。如前文所述，基于GAN的视频超分辨率网络的研究才刚开始，有很多方案可以尝试。non-local处理多帧特征是一种极佳思路，PFNL和MuCAN都展现了不俗的性能，但PFNL的参数量仅为3. 0 M，重建720 × 1 280像素视频帧的时间却需要295 ms，参数量相差不多的基于循环神经网络的非图像配准方法RRN仅需45 ms，基于可变形卷积的图像配准方法EDVR-M参数量达到3. 3 M，但重建720 × 1 280像素视频帧运行时间也仅需118 ms。non-local的计算复杂度难以忽视，但对于性能的提升却是显著的，前文介绍的YouKu 2020视频超分挑战赛冠军VESR-Net也是在可变形卷积的基础上加入了non-local模块。non-local与图像配准相结合的方法比直接使用non-local可能更适用于视频超分辨率任务。

基于图像配准的方法与基于非图像配准的方法正在并行发展。非图像配准的方法希望能用较少参数量、较小的运算负荷，依靠循环网络、3D卷积等方式隐式地利用帧间差异实现视频重构。2020年提出的RRN、RSDN、RLSP都展现出在实时性重建出高质量、高分辨率视频上的优势。在此基础上，如何更好地提升重建视频质量将是此类方法未来的发展方向。基于图像配准的方法已经在×2，×3，×4的放大因子上表示出了良好的性能，2020年的视频超分辨挑战赛都将目光放到16倍视频超分上，且在挑战赛中名列前茅的视频超分辨率方法都为利用可变形卷积的基于图像配准方法。面对更大放大因子的挑战，如何获取更丰富的细节信息是基于图像配准方法的主要研究方向。可以从如何引入更为高效的注意力机制和设计更准确的图像配准方法等方面考虑，进一步提升重建视频质量。

功能不同的两类方法在许多方面可以互相借鉴共同发展。例如，非图像配准方法PFNL引入的non-local使注意力机制也成为图像配准方法中的热点模块。循环神经网络在图像配准方法的应用BasicVSR++和在非图像配准方法的应用OVSR都实现了非常好的重建效果。

5 未来的发展与挑战

基于深度学习的视频超分辨率技术已经取得了重大进展，特别是在一些公共数据集上实现了远好于传统非深度学习方法的效果。不同模型的特点与局限性已经在前文分别讨论过，本节将概括分析目前亟待解决的问题，并对接下来的研究趋势进行展望，希望对相关研究人员有一些启发。

5.1 网络训练及优化

为了追求好的视频重构效果，恢复丰富明显的细节，损失函数作为高分辨率视频与重构视频之间的约束，它的选取特别重要。从表16可以看出，损失函数的选取比较单调，且没有确切的理论依据。与此相似，评价指标大多为PSNR和SSIM，但是基于GAN的视频超分辨率网络使用此类指标衡量则并不合理，有些视频的重建视觉效果很好，但在指标上却获得不尽如人意的质量分数，有些主观指标存在评价误差，且耗时耗力。因此，评价指标的设计具有很大的研究价值。

在网络优化方面，权重参数的初始化、激活函数的选取以及数据增强策略都是需要研究的重要问题。Liu等人(2021)使用NAS(Kim等，2019a)选取特定的超参数和寻找最优网络结构，进一步优化RRN(Isobe等，2020c)网络，获得了MAI 2021挑战赛的第2名。在CVPR 2021中，Xiao等人(2021)基于视频超分辨率网络的特性，设计了时空蒸馏的策略，使轻量级网络有比大型复杂网络更好的性能。这些都为未来的视频超分辨率研究提供了新的思考方向。

5.2 超高分辨率视频的重建

现有的基于深度学习的视频超分辨率网络在训练与测试时使用的数据集主要集中于Vimeo90K(448×256像素)和Vid4(720×480像素)等低分辨率视频序列。受数据集分辨率和网络结构限制，已有方法难以适应高分辨率场景的需求，如2 K(2 048×1 080像素)、QHD(2 560×1 440像素)、UHD(3 840×2 160像素)和8 K(7 680×4 320像素)等。面对需要重建出更加复杂纹理和更清晰细节的挑战，设计和实现一个轻量级的、高效的视频超分辨率算法将极具现实意义。

5.3 压缩视频的超分辨率

传统视频超分辨率网络使用未压缩的数据来重建高质量的视频，在对高度压缩的输入视频进行超分辨率时往往会产生严重的伪影。在压缩帧上训练现有最先进的VSR模型，可以使模型具有一定处理压缩信息的能力。但如果不对网络模块的设计进行具体更改，这些训练数据甚至可能会损害整体性能。周亮和朱秀昌(2006)运用Bayesian估计理论，在最大后验概率准则下对视频进行了有效的压缩。Liu等人(2021)设计了一种压缩感知超分辨率模型在一定程度上恢复了由于视频压缩导致的丢失信息。Chen等人(2021)不仅在网络上进行了设计，而且利用压缩先验的引导更加细粒度地挖掘特征信息，从而获得更好的超分辨率性能。目前，基于压缩视频的超分辨率工作较少，在未来这类方法将会获得更多的关注。

5.4 多降质联合视频缩放

视频超分辨率的数据集使用的是人工下采样的低分辨率视频序列，这些与自然情况下退化的视频相差较大。降质模型太简单，无法描述现实场景中运动变形、光学模糊、降采样和噪声污染等多种复杂降质因素的影响。仿真HR/LR数据集的不准确性影响了视频超分辨率算法在现实应用中的表现。Yang等人(2021)尝试利用多相机成像和Doubletake软件生成真实数据集，提高模型在真实场景应用时重建视频的视觉质量。但两个传感器采集的高低分辨率视频由于焦距、感光以及ISP算法等方面的差异不可避免地会存在亮度、色彩的不一致。

多降质联合(复杂降质)的视频缩放可以同时学习超分辨率网络和下采样网络，以适应真实的低分辨率视频。Huang等人(2021)设计了多输入多输出视频缩放网络(multi-input multi-output video rescaling network，MIMO-VRN)。MIMO-VRN提出了第1个基于可逆耦合结构的视频上采样和降采样联合优化方案。如何更好地同时学习成对缩放过程以应对复杂退化过程的低分辨率视频序列有待探讨。

5.5　自监督学习视频超分辨率

有监督学习依赖大量真实成对的LR和HR视频帧。然而，这样的配对数据集在现实中很少见，在实践中获取真实数据集花费巨大。自监督学习可以只根据输入的低分辨率视频进行超分辨率网络的训练和推断，而无需其他监督信息。因此，探索基于自监督学习的视频超分辨率技术具有重要的研究意义。

5.6　特定应用场景视频超分辨率

视频超分辨率技术的研究已经有较强的普适性。但在面对特定应用场景时，需要将相关领域的先验知识与视频超分辨率算法结合。在视频监控、智能运输和无人驾驶等领域，视频超分辨率网络需要从特定的区域，如人脸、指纹、车牌、障碍物和车道线等获取清晰的细节信息。张岩等人(2016)针对无人机摄像、照相数据的特点，提出一种无人机侦察视频超分辨率重建方法。利用特定领域的先验知识指导视频超分辨率研究，可以提高生成的视频质量，帮助改进其他计算机视觉任务。

5.7　任意比例因子视频超分辨率

众所周知，大多数现有的视频超分辨率方法仅考虑某些整数比例因子(×2，×3，×4)的超分辨率。而用于超分辨率的比例因子应该可以是任何正数，并且不应将其固定为某些特定整数。目前的方法都将不同比例因子的超分辨率视为独立任务。但如果为每个比例因子训练一个特定的模型，则不可能存储所有对应比例因子的模型，并且计算效率低下。因此，可以设计实现任意比例因子的视频超分辨率模型。

5.8　时空视频超分辨率

目前，基于深度学习的空间视频超分辨率技术已经很好地改善视频视觉效果。但是，视频的质量不仅取决于分辨率，也受帧率的影响。将时间域的视频插帧技术与空间视频超分辨率技术相结合，设计同时利用时空信息的超分辨率网络(Xiang等，2020；Xu等，2021；Dutta等，2021)，生成高分辨率、高帧率的视频将会是未来研究的重点工作。

5.9　辅助任务正则视频超分辨率

现有大多数方法往往直接利用视频帧的内容信息进行视频超分辨率的研究。在解决其他计算机视觉任务时，Bao等人(2019)利用深度信息作为上下文特征帮助视频插帧。Nazeri等人(2019)利用图像边缘特征和结构信息辅助单图超分辨率。这种利用额外信息辅助目标任务往往能获取非凡的效果。CVPR 2021中，Jing等人(2021)将事件相机中的高频数据作为视频超分辨率网络RBPN(Haris等，2019)的额外输入，最终提高了视频帧的质量。而在视频超分辨率领域，利用辅助信息的研究较少。因此，在未来研究过程中，可以更多地利用深度估计、边缘检测等方法获取视频帧中的深层特征，进而提升视频超分辨率效果。

6　结　语

本文首先介绍了视频超分辨率的意义以及研究背景，对其评价指标以及数据集进行了说明。随后，总结了近几年国内外对基于深度学习的视频超分辨率算法的研究，并将其分为基于图像配准的方法和非图像配准的方法，按网络特性进行分类概述，并设计实验在几个标准数据集中对网络进行评估，最后对目前的挑战进行分析。总体而言，本文对视频超分辨率领域的既有研究做出了进一步的总结和方向探索，期望能够启发读者在未来设计更高效的视频超分辨率网络。基于深度学习的视频超分辨率技术是目前十分活跃的研究方向，相信未来也会有更多的技术和研究成果出现，综述的内容也会不断地更新和扩展。

参考文献(References)

Ahmadi A and Patras I. 2016. Unsupervised convolutional neural networks for motion estimation//Proceedings of 2016 IEEE International Conference on Image Processing (ICIP). Phoenix, USA: IEEE: 1629-1633 [DOI: 10.1109/ICIP.2016.7532634]

Banham M R and Katsaggelos A K. 1997. Digital image restoration. IEEE Signal Processing Magazine, 14(2): 24-41 [DOI: 10.1109/79.581363]

Bao W B, Lai W S, Ma C, Zhang X Y, Gao Z Y and Yang M H. 2019a. Depth-aware video frame interpolation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3698-3707 [DOI: 10.1109/CVPR.2019.00382]

Bao W B, Lai W S, Zhang X Y, Gao Z Y and Yang M H. 2019b. Memc-net: motion estimation and motion compensation driven neural network for video interpolation and enhancement. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(3): 933-948

[DOI: 10.1109/TPAMI.2019.2941941]

Bare B, Yan B, Ma C X and Li K. 2019. Real-time video super-resolution via motion convolution kernel estimation. Neurocomputing, 367: 236-245 [DOI: 10.1016/j.neucom.2019.07.089]

Bertasius G, Torresani L and Shi J B. 2018. Object detection in video with spatiotemporal sampling networks//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 342-357 [DOI: 10.1007/978-3-030-01258-8_21]

Bouguet J Y. 2001. Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. Intel Corporation, 5(4): 1-10

Brox T, Bruhn A, Papenberg N and Weickert J. 2004. High accuracy optical flow estimation based on a theory for warping//Proceedings of the 8th European Conference on Computer Vision (ECCV). Prague, Czech Republic: Springer: 25-36 [DOI: 10.1007/978-3-540-24673-2_3]

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]

Cao J Z, Li Y W, Zhang K and van Gool L. 2021. Video super-resolution transformer [EB/OL]. [2022-02-08]. https://arxiv.org/pdf/2106.06847.pdf

Chan K C K, Wang X T, Yu K, Dong C and Loy C C. 2021a. Basicvsr: the search for essential components in video super-resolution and beyond//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 4945-4954 [DOI: 10.1109/CVPR46437.2021.00491]

Chan K C K, Zhou S C, Xu X Y and Loy C C. 2021b. Basicvsr++: improving video super-resolution with enhanced propagation and alignment//Proceedings of 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR). New Orleans, USA: IEEE: 5972-5981 [DOI: 10.1109/CVPR52688.2022.00588]

Chen J L, Tan X, Shan C W, Liu S and Chen Z B. 2020. VESR-Net: the winning solution to YouKu video enhancement and super-resolution challenge [EB/OL]. [2022-02-08]. https://arxiv.org/pdf/2003.02115.pdf

Chen P L, Yang W H, Wang M, Sun L, Hu K K and Wang S Q. 2021. Compressed domain deep video super-resolution. IEEE Transactions on Image Processing, 30: 7156-7169 [DOI: 10.1109/TIP.2021.3101826]

Chen Y, Tai Y, Liu X M, Shen C H and Yang J. 2018. FSRNet: end-to-end learning face super-resolution with facial priors//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 2492-2501 [DOI: 10.1109/CVPR.2018.00264]

Cheng D Q, Guo X, Chen L L, Kou Q Q, Zhao K and Gao R. 2021. Image super-resolution reconstruction from multi-channel recursive residual network. Journal of Image and Graphics, 26(3): 605-618 (程德强, 郭昕, 陈亮亮, 寇旗旗, 赵凯, 高蕊. 2021. 多通道递归残差网络的图像超分辨率重建. 中国图象图形学报, 26(3): 605-618) [DOI: 10.11834/jig.200108]

Cheng M H, Lin N W, Hwang K S and Jeng J H. 2012. Fast video super-resolution using artificial neural networks//Proceedings of the 8th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP). Poznan, Poland: IEEE: 1-4 [DOI: 10.1109/CSNDSP.2012.6292646]

Chu M Y, Xie Y, Mayer J, Leal-Taixé L and Thuerey N. 2020. Learning temporal coherence via self-supervision for gan-based video generation. ACM Transactions on Graphics, 39(4): #75 [DOI: 10.1145/3386569.3392457]

Cui Z, Chang H, Shan S G, Zhong B N and Chen X L. 2014. Deep network cascade for image super-resolution//Proceedings of the 13th European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer: 49-64 [DOI: 10.1007/978-3-319-10602-1_4]

Dai J F, Qi H Z, Xiong Y W, Li Y, 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]

Dong C, Loy C C, He K M and Tang X O. 2014. Learning a deep convolutional network for image super-resolution//Proceedings of the 13th European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer: 184-199 [DOI: 10.1007/978-3-319-10593-2_13]

Dong C, Loy C C, He K M and Tang X O. 2016. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(2): 295-307 [DOI: 10.1109/TPAMI.2015.2439281]

Dosovitskiy A, Fischer P, Ilg E, Häusser P, Hazirbas C, Golkov V, van der Smagt P, Cremers D and Brox T. 2015. FlowNet: learning optical flow with convolutional networks//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE: 2758-2766 [DOI: 10.1109/ICCV.2015.316]

Drulea M and Nedevschi S. 2011. Total variation regularization of local-global optical flow//Proceedings of the 14th International IEEE Conference on Intelligent Transportation Systems (ITSC). Washington, USA: IEEE: 318-323 [DOI: 10.1109/ITSC.2011.6082986]

Dutta S, Shah N A and Mittal A. 2021. Efficient space-time video super resolution using low-resolution flow and mask upsampling//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Nashville, USA: IEEE: 314-323 [DOI: 10.1109/CVPRW53098.2021.00041]

Ebadi S E, Ones V G and Izquierdo E. 2017. Uhd video super-resolution using low-rank and sparse decomposition//Proceedings of 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). Venice, Italy: IEEE: 1889-1897 [DOI: 10.1109/

ICCVW.2017.223］

Farnebäck G. 2003. Two-frame motion estimation based on polynomial expansion//Proceedings of the 13th Scandinavian Conference on Image Analysis. Halmstad, Sweden: Springer: 363-370［DOI: 10.1007/3-540-45103-X_50］

Fuoli D, Gu S H and Timofte R. 2019. Efficient video super-resolution through recurrent latent space propagation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul, Korea (South): IEEE: 3476-3485［DOI: 10.1109/ICCVW.2019.00431］

Fuoli D, Huang Z W, Gu S H, Timofte R, Raventos A, Esfandiari A, Karout S, Xu X, Li X, Xiong X, Wang J G, Michelini P N, Zhang W H, Zhang D Y, Zhu H W, Xia D, Chen H Y, Gu J J, Zhang Z, Zhao T T, Zhao S S, Akita K, Ukita N, Hrishikesh P S, Puthussery D and Jiji C V. 2020. AIM 2020 challenge on video extreme super-resolution: methods and results//Proceedings of 2020 European Conference on Computer Vision (ECCV). Glasgow, UK: Springer: 57-81［DOI: 10.1007/978-3-030-66823-5_4］

Ganin Y, Kononenko D, Sungatullina D and Lempitsky V. 2016. DeepWarp: photorealistic image resynthesis for gaze manipulation//Proceedings of the 14th European Conference on Computer Vision (ECCV). Amsterdam, the Netherlands: Springer: 311-326［DOI: 10.1007/978-3-319-46475-6_20］

Gao H, Zhu X Z, Lin S and Dai J F. 2019. Deformable kernels: adapting effective receptive fields for object deformation［EB/OL］.［2022-02-08］. https://arxiv.org/pdf/1910.02940v1.pdf

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial networks. Communications of the ACM, 63(11): 139-144［DOI: 10.1145/3422622］

Graves A, Fernández S and Schmidhuber J. 2005. Bidirectional LSTM networks for improved phoneme classification and recognition//Proceedings of the 15th International Conference on Artificial Neural Networks: Formal Models and Their Applications. Warsaw, Poland: Springer: 799-804［DOI: 10.1007/11550907_126］

Gunturk B K, Batur A U, Altunbasak Y, Hayes M H and Mersereau R M. 2003. Eigenface-domain super-resolution for face recognition. IEEE Transactions on Image Processing, 12(5): 597-606［DOI: 10.1109/TIP.2003.811513］

Guo J and Chao H Y. 2017. Building an end-to-end spatial-temporal convolutional network for video super-resolution. Proceedings of the AAAI Conference on Artificial Intelligence, 31(1): 4053-4060［DOI: 10.1609/aaai.v31i1.11228］

Handa A, Bloesch M, Pătrăucean V, Stent S, McCormac J and Davison A. 2016. Gvnn: neural network library for geometric computer vision//Proceedings of 2016 European Conference on Computer Vision (ECCV). Amsterdam, the Netherlands: Springer: 67-82［DOI: 10.1007/978-3-319-49409-8_9］

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］

Haris M, Shakhnarovich G and Ukita N. 2020. Space-time-aware multi-resolution video enhancement//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Atlanta, USA: IEEE: 2856-2865［DOI: 10.1109/CVPR42600.2020.00293］

Harris J L. 1964. Diffraction and resolving power. Journal of the Optical Society of America, 54(7): 931-936［DOI: 10.1364/josa.54.000931］

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］

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 770-778［DOI: 10.1109/CVPR.2016.90］

He X H, Wu Y Y, Chen W L and Qing L B. 2011. A survey of video super-resolution reconstruction technology. Information and Electronic Engineering, 9(1): 1-6 (何小海, 吴媛媛, 陈为龙, 卿粼波. 2011. 视频超分辨率重建技术综述. 信息与电子工程, 9(1): 1-6)［DOI: 10.3969/j.issn.1672-2892.2011.01.001］

Hochreiter S and Schmidhuber J. 1997. Long short-term memory. Neural Computation, 9(8): 1735-1780［DOI: 10.1162/neco.1997.9.8.1735］

Huang G, Liu Z, van der Maaten L and Weinberger K Q. 2017. Densely connected convolutional networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 2261-2269［DOI: 10.1109/CVPR.2017.243］

Huang T S and Tsai R Y. 1984. Multiframe image restoration and registration//Advances in Computer Vision and Image Processing. Greenwich, UK: JAI Press: 317-339

Huang Y, Wang W and Wang L. 2015. Bidirectional recurrent convolutional networks for multi-frame super-resolution［EB/OL］.［2022-02-08］. http://cognn.com/papers/24%20NIPS%202015%20Yan%20bidirectional-recurrent-convolutional-networks-for-multi-frame-super-resolution-Paper.pdf

Huang Y, Wang W and Wang L. 2018. Video super-resolution via bidirectional recurrent convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4): 1015-1028［DOI: 10.1109/TPAMI.2017.2701380］

Huang Y C, Chen Y H, Lu C Y, Wang H P, Peng W H and Huang C

C. 2021. Video rescaling networks with joint optimization strategies for downscaling and upscaling//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 3526-3535 [DOI: 10.1109/CVPR46437.2021.00353]

Hui Z, Li J, Gao X B and Wang X M. 2021. Progressive perception-oriented network for single image super-resolution. Information Sciences, 546: 769-786 [DOI: 10.1016/j.ins.2020.08.114]

Ilg E, Mayer N, Saikia T, Keuper M, Dosovitskiy A and Brox T. 2017. FlowNet 2.0: evolution of optical flow estimation with deep networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 1647-1655 [DOI: 10.1109/CVPR.2017.179]

Isobe T, Jia X, Gu S H, Li S J, Wang S J and Tian Q. 2020a. Video super-resolution with recurrent structure-detail network//Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK: Springer: 645-660 [DOI: 10.1007/978-3-030-58610-2_38]

Isobe T, Li S J, Jia X, Yuan S X, Slabaugh G, Xu C J, Li Y L, Wang S J and Tian Q. 2020b. Video super-resolution with temporal group attention//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 8005-8014 [DOI: 10.1109/CVPR42600.2020.00803]

Isobe T, Zhu F, Jia X and Wang S J. 2020c. Revisiting temporal modeling for video super-resolution [EB/OL]. [2022-02-08]. https://arxiv.org/pdf/2008.05765.pdf

Jaderberg M, Simonyan K, Zisserman A and Kavukcuoglu K. 2016. Spatial transformer networks [EB/OL]. [2022-02-08]. https://arxiv.org/pdf/1506.02025.pdf

Ji S W, Xu W, Yang M and Yu K. 2013. 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1): 221-231 [DOI: 10.1109/TPAMI.2012.59]

Jing Y C, Yang Y D, Wang X C, Song M L and Tao D C. 2021. Turning frequency to resolution: video super-resolution via event cameras//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 7768-7777 [DOI: 10.1109/CVPR46437.2021.00768]

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]

Kalarot R and Porikli F. 2019. MultiBoot Vsr: multi-stage multi-reference bootstrapping for video super-resolution//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, USA: IEEE: 2060-2069 [DOI: 10.1109/CVPRW.2019.00258]

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]

Kim H, Hong S, Han B, Myeong H and Lee K M. 2019a. Fine-grained neural architecture search [EB/OL]. [2022-02-08]. https://arxiv.org/pdf/1911.07478.pdf

Kim J, Lee J K and Lee K M. 2016. Accurate image super-resolution using very deep convolutional networks//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 1646-1654 [DOI: 10.1109/CVPR.2016.182]

Kim S Y, Lim J, Na T and Kim M. 2019b. Video super-resolution based on 3D-CNNS with consideration of scene change//Proceedings of 2019 IEEE International Conference on Image Processing (ICIP). Taipei, China: IEEE: 2831-2835 [DOI: 10.1109/ICIP.2019.8803297]

Kim T H, Sajjadi M S M, Hirsch M and Schölkopf B. 2018. Spatio-temporal transformer network for video restoration//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 111-127 [DOI: 10.1007/978-3-030-01219-9_7]

Lai W S, Huang J B, Ahuja N and Yang M H. 2017. Deep Laplacian pyramid networks for fast and accurate super-resolution//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 5835-5843 [DOI: 10.1109/CVPR.2017.618]

Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z H and Shi W Z. 2017. Photo-realistic single image super-resolution using a generative adversarial network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 105-114 [DOI: 10.1109/CVPR.2017.19]

Lertrattanapanich S and Bose N K. 1999. Latest results on high-resolution reconstruction from video sequences [EB/OL]. [2022-02-08]. https://www.semanticscholar.org/paper/Latest-Results-on-High-Resolution-ReconstructionLertrattanapanich/bd8bc32eaf0ffd502d008c36f2c1d870e12ea238

Li D Y, Liu Y and Wang Z F. 2019a. Video super-resolution using non-simultaneous fully recurrent convolutional network. IEEE Transactions on Image Processing, 28(3): 1342-1355 [DOI: 10.1109/TIP.2018.2877334]

Li D Y and Wang Z F. 2017. Video superresolution via motion compensation and deep residual learning. IEEE Transactions on Computational Imaging, 3(4): 749-762 [DOI: 10.1109/TCI.2017.2671360]

Li K, Bare B, Yan B, Feng B L and Yao C F. 2018. Face hallucination based on key parts enhancement//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing

(ICASSP). Calgary, Canada: IEEE: 1378-1382 [DOI: 10.1109/ICASSP.2018.8462170]

Li S, He F X, Du B, Zhang L F, Xu Y H and Tao D C. 2019b. Fast spatio-temporal residual network for video super-resolution//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 10514-10523 [DOI: 10.1109/CVPR.2019.01077]

Li S L, Feng C L, Yu K, Liu X, Jiang X and Zhao D Z. 2022. Critical review of human cardiac magnetic resonance image super resolution reconstruction based on deep learning method. Journal of Image and Graphics, 27(3): 704-721 (李书林, 冯朝路, 于鲲, 刘鑫, 江鑫, 赵大哲. 2022. 基于深度学习的心脏磁共振影像超分辨率前沿进展. 中国图象图形学报, 27(3): 704-721) [DOI: 10.11834/jig.210150]

Li W B, Tao X, Guo T A, Qi L, Lu J B and Jia J Y. 2020. MuCAN: multi-correspondence aggregation network for video super-resolution//Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK: Springer: 335-351 [DOI: 10.1007/978-3-030-58607-2_20]

Li Y, Jin P, Yang F, Liu C, Yang M H and Milanfar P. 2021. COMISR: compression-informed video super-resolution//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 2543-2552 [DOI: 10.1109/ICCV48922.2021.00254]

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. 2011. A Bayesian approach to adaptive video super resolution//Proceedings of 2011 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Colorado Springs, USA: IEEE: 209-216 [DOI: 10.1109/CVPR.2011.5995614]

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]

Liu H Y, Ruan Z B, Zhao P, Dong C, Shang F H, Liu Y Y, Yang L L and Timofte R. 2022. Video super resolution based on deep learning: a comprehensive survey. Artificial Intelligence Review, 55(8): 5981-6035 [DOI: 10.1007/s10462-022-10147-y]

Liu S L, Zheng C J, Lu K D, Gao S, Wang N, Wang B F, Zhang D K, Zhang X F and Xu T Y. 2021. EVSRNet: efficient video super-resolution with neural architecture search//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Nashville, USA: IEEE: 2480-2485 [DOI: 10.1109/CVPRW53098.2021.00281]

Lucas A, Lopez-Tapia S, Molina R and Katsaggelos A K. 2019. Generative adversarial networks and perceptual losses for video super-resolution. IEEE Transactions on Image Processing, 28(7): 3312-3327 [DOI: 10.1109/TIP.2019.2895768]

Lucas B D and Kanade T. 1981. An iterative image registration technique with an application to stereo vision//Proceedings of the 7th international joint conference on Artificial intelligence. Vancouver BC, Canada: Morgan Kaufmann Publishers Inc: 674-679

Mao X J, Shen C H and Yang Y B. 2016. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc.: 2810-2818 [DOI: 10.5555/3157382.3157412]

Nah S, Baik S, Hong S, Moon G, Son S, Timofte R and Lee K M. 2019a. NTIRE 2019 challenge on video deblurring and super-resolution: dataset and study//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, USA: IEEE: 1996-2005 [DOI: 10.1109/CVPRW.2019.00251]

Nah S, Timofte R, Gu S H, Baik S, Hong S, Moon G, Son S, Lee K M, Wang X T, Chan K C K, Yu K, Dong C, Loy C C, Fan Y C, Yu J H, Liu D, Huang T S, Liu X, Li C, He D L, Ding Y K, Wen S L, Porikli F, Kalarot R, Haris M, Shakhnarovich G, Ukita N, Yi P, Wang Z Y, Jiang K, Jiang J J, Ma J Y, Dong H, Zhang X Y, Hu Z, Kim K, Kang D U, Chun S Y, Purohit K, Rajagopalan A N, Tian Y P, Zhang Y L, Fu Y, Xu C L, Tekalp A M, Yilmaz M A, Korkmaz C, Sharma M, Makwana M, Badhwar A, Singh A P, Upadhyay A, Mukhopadhyay R, Shukla A, Khanna D, Mandal A S, Chaudhury S, Miao S, Zhu Y X and Huo X. 2019b. NTIRE 2019 challenge on video super-resolution: methods and results//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, USA: IEEE: 1985-1995 [DOI: 10.1109/CVPRW.2019.00250]

Nazeri K, Thasarathan H and Ebrahimi M. 2019. Edge-informed single image super-resolution//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop (CVPRW). Seoul, Korea (South): IEEE: 3275-3284 [DOI: 10.1109/ICCVW.2019.00409]

Niklaus S, Mai L and Liu F. 2017a. Video frame interpolation via adaptive separable convolution//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 261-270 [DOI: 10.1109/ICCV.2017.37]

Niklaus S, Mai L and Liu F. 2017b. Video frame interpolation via adaptive convolution//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2270-2279 [DOI: 10.1109/CVPR.2017.244]

Patraucean V, Handa A and Cipolla R. 2016. Spatio-temporal video

autoencoder with differentiable memory [EB/OL]. [2022-02-08]. https://arxiv.org/pdf/1511.06309.pdf

Protter M, Elad M, Takeda H and Milanfar P. 2009. Generalizing the nonlocal-means to super-resolution reconstruction. IEEE Transactions on Image Processing, 18(1): 36-51 [DOI: 10.1109/TIP.2008.2008067]

Ranjan A and Black M J. 2017. Optical flow estimation using a spatial pyramid network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 2720-2729 [DOI: 10.1109/CVPR.2017.291]

Revaud J, Weinzaepfel P, Harchaoui Z and Schmid C. 2015. EpicFlow: edge-preserving interpolation of correspondences for optical flow//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 1164-1172 [DOI: 10.1109/CVPR.2015.7298720]

Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

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]

Schuster M and Paliwal K K. 1997. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11): 2673-2681 [DOI: 10.1109/78.650093]

Seshadrinathan K and Bovik A C. 2010. Motion tuned spatio-temporal quality assessment of natural videos. IEEE Transactions on Image Processing, 19(2): 335-350 [DOI: 10.1109/TIP.2009.2034992]

Shahar O, Faktor A and Irani M. 2011. Space-time super-resolution from a single video//Proceedings of 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Colorado Springs, USA: IEEE: 3353-3360 [DOI: 10.1109/CVPR.2011.5995360]

Sheikh H R, Sabir M F and Bovik A C. 2006. A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Transactions on Image Processing, 15(11): 3440-3451 [DOI: 10.1109/TIP.2006.881959]

Shi W Z, Caballero J, Huszár F, Totz, Aitken A P, Bishop R, Rueckert D and Wang Z. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 1874-1883 [DOI: 10.1109/CVPR.2016.207]

Shi X J, Chen Z R, Wang H, Yeung D Y, Wong W K and Woo W C. 2015. Convolutional LSTM network: a machine learning approach for precipitation nowcasting//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 802-810

Singh A and Singh J. 2020. Survey on single image based super-resolutio — implementation challenges and solutions. Multimedia Tools and Applications, 79(3): 1641-1672 [DOI: 10.1007/s11042-019-08254-0]

Sun D Q, Yang X D, Liu M Y and Kautz J. 2018a. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE: 8934-8943 [DOI: 10.1109/CVPR.2018.00931]

Sun X, Xiao B, Wei F Y, Liang S and Wei Y C. 2018b. Integral human pose regression//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 536-553 [DOI: 10.1007/978-3-030-01231-1_33]

Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A. 2015. Going deeper with convolutions//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 1-9 [DOI: 10.1109/CVPR.2015.7298594]

Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z. 2016. Rethinking the inception architecture for computer vision//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 2818-2826 [DOI: 10.1109/CVPR.2016.308]

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 (CVPR). Seattle, USA: IEEE: 3357-3366 [DOI: 10.1109/CVPR42600.2020.00342]

Toderici G, O'Malley S M, Hwang S J, Vincent D, Minnen D, Baluja S, Covell M and Sukthankar R. 2016. Variable rate image compression with recurrent neural networks [EB/OL]. [2022-02-08]. https://arxiv.org/pdf/1511.06085v5.pdf

Tong T, Li G, Liu X J and Gao Q Q. 2017. Image super-resolution using dense skip connections//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 4809-4817 [DOI: 10.1109/ICCV.2017.514]

Tran D, Bourdev L, Fergus R, Torresani L and Paluri M. 2015. Learning spatiotemporal features with 3D convolutional networks//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE: 4489-4497 [DOI: 10.1109/ICCV.2015.510]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need [EB/OL]. [2022-02-08]. https://arxiv.org/pdf/1706.03762.pdf

Wang H, Su D W, Liu C C, Jin L C, Sun X F and Peng X Y. 2019a.

Deformable non-local network for video super-resolution. IEEE Access, 7: 177734-177744 [DOI: 10.1109/ACCESS.2019.2958030]

Wang L G, Guo Y L, Lin Z P, Deng X P and An W. 2018a. Learning for video super-resolution through HR optical flow estimation//Proceedings of the 14th Asian Conference on Computer Vision. Perth, Australia: Springer: 514-529 [DOI: 10.1007/978-3-030-20887-5_32]

Wang X L, Girshick R, Gupta A and He K M. 2018b. Non-local neural networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE: 7794-7803 [DOI: 10.1109/CVPR.2018.00813]

Wang X T, Chan K C K, Yu K, Dong C and Loy C C. 2019b. 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. 2018c. Recovering realistic texture in image super-resolution by deep spatial feature transform//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 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]

Wang Z H, Chen J and Hoi S C H. 2021. Deep learning for image super-resolution: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(10): 3365-3387 [DOI: 10.1109/TPAMI.2020.2982166]

Wang Z W, Liu D, Yang J C, Han W and Huang T. 2015a. Deep networks for image super-resolution with sparse prior//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE: 370-378 [DOI: 10.1109/ICCV.2015.50]

Wang Z Y, Yang Y Z, Wang Z W, Chang S Y, Han W, Yang J C and Huang T. 2015b. Self-tuned deep super resolution//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Boston, USA: IEEE: 1-8 [DOI: 10.1109/CVPRW.2015.7301266]

Wang Z Y, Yi P, Jiang K, Jiang J J, Han Z, Lu T and Ma J Y. 2019c. Multi-memory convolutional neural network for video super-resolution. IEEE Transactions on Image Processing, 28(5): 2530-2544 [DOI: 10.1109/TIP.2018.2887017]

Wolf S, Pinson M H. 2011, Video quality model for variable frame delay (VQM-VFD). [EB/OL]. [2022-02-08]. https://last.hit.bme.hu/download/vidtechlab/fcc/literature/video/ntia_tm-11-482.pdf

Wu Y and Fan G H. 2017. Survey of super-resolution reconstruction techniques for video sequences. Computer Engineering and Software, 38(4): 154-160 (吴洋, 樊桂花. 2017. 视频序列超分辨率重构技术综述. 软件, 38(4): 154-160) [DOI: 10.3969/j.issn.1003-6970.2017.04.030]

Xiang X Y, Tian Y P, Zhang Y L, Fu Y, Allebach J P and Xu C L. 2020. Zooming Slow-Mo: fast and accurate one-stage space-time video super-resolution//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 3367-3376 [DOI: 10.1109/CVPR42600.2020.00343]

Xiao Z Y, Fu X Y, Huang J, Cheng Z and Xiong Z W. 2021. Space-time distillation for video super-resolution//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 2113-2122 [DOI: 10.1109/CVPR46437.2021.00215]

Xu G, Xu J, Li Z, Wang L, Sun X and Cheng M M. 2021. Temporal modulation network for controllable space-time video super-resolution//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 6384-6393 [DOI: 10.1109/CVPR46437.2021.00632]

Xu J, Chae Y, Stenger B and Datta A. 2018. Dense bynet: residual dense network for image super resolution//Proceedings of the 25th IEEE International Conference on Image Processing (ICIP). Athens, Greece: IEEE: 71-75 [DOI: 10.1109/ICIP.2018.8451696]

Xu L, Jia J Y and Matsushita Y. 2012. Motion detail preserving optical flow estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(9): 1744-1757 [DOI: 10.1109/TPAMI.2011.236]

Xu X, Xiong X, Wang J G and Li X. 2020. Deformable kernel convolutional network for video extreme super-resolution//Proceedings of 2020 European Conference on Computer Vision (ECCV). Glasgow, UK: Springer: 82-98 [DOI: 10.1007/978-3-030-66823-5_5]

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]

Yang W M, Zhang X C, Tian Y P, Wang W, Xue J H and Liao Q M. 2019. Deep learning for single image super-resolution: a brief review. IEEE Transactions on Multimedia, 21(12): 3106-3121 [DOI: 10.1109/TMM.2019.2919431]

Yang X, Xiang W M, Zeng H and Zhang L. 2021. Real-world video super-resolution: a benchmark dataset and a decomposition based learning scheme//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 4761-4770 [DOI: 10.1109/ICCV48922.2021.00474]

Yi P, Wang Z Y, Jiang K, Jiang J J, Lu T, Tian X and Ma J Y. 2021. Omniscient video super-resolution//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 4429-4438 [DOI: 10.1109/ICCV48922.2021.00439]

Yi P, Wang Z Y, Jiang K, Jiang J J and Ma J Y. 2019. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 3106-3115 [DOI: 10.1109/ICCV.2019.00320]

Ying X Y, Wang L G, Wang Y Q, Sheng W D, An W and Guo Y L. 2020. Deformable 3D convolution for video super-resolution. IEEE Signal Processing Letters, 27: 1500-1504 [DOI: 10.1109/LSP.2020.3013518]

Yu F and Koltun V. 2016. Multi-scale context aggregation by dilated convolutions [EB/OL]. [2022-02-08]. https://arxiv.org/pdf/1511.07122v2.pdf

Zhang L P, Zhang H Y, Shen H F and Li P X. 2010. A super-resolution reconstruction algorithm for surveillance images. Signal Processing, 90(3): 848-859 [DOI: 10.1016/j.sigpro.2009.09.002]

Zhang R, Isola P, Efros A A, Shechtman E and Wang O. 2018a. The unreasonable effectiveness of deep features as a perceptual metric//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE: 586-595 [DOI: 10.1109/CVPR.2018.00068]

Zhang Y, Li J Z, Li D L and Du Y L. 2016. Super-resolution reconstruction for UAV video. Journal of Image and Graphics, 21(7): 967-976 (张岩, 李建增, 李德良, 杜玉龙. 2016. 无人机侦察视频超分辨率重建方法. 中国图象图形学报, 21(7): 967-976) [DOI: 10.11834/jig.20160715]

Zhang Y L, Gan Z L and Zhu X C. 2013. Video super-resolution method based on similarity constraints. Journal of Image and Graphics, 18(7): 761-767 (张义轮, 干宗良, 朱秀昌. 2013. 相似性约束的视频超分辨率重建. 中国图象图形学报, 18(7): 761-767) [DOI: 10.11834/jig.20130712]

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

Zhou B, Li C H and Chen W. 2021. Region-level channel attention for single image super-resolution combining high frequency loss. Journal of Image and Graphics, 26(12): 2836-2847 (周波, 李成华, 陈伟. 2021. 区域级通道注意力融合高频损失的图像超分辨率重建. 中国图象图形学报, 26(12): 2836-2847) [DOI: 10.11834/jig.200582]

Zhou L and Zhu X C. 2006. Algorithm of compressed video super-resolution restoration based on bayesian theory. Journal of Image and Graphics, 11(5): 730-735 (周亮, 朱秀昌. 2006. 基于Bayesian理论的压缩视频超分辨率重构算法. 中国图象图形学报, 11(5): 730-735) [DOI: 10.11834/jig.200605121]

Zhu X B, Li Z Z, Zhang X Y, Li C S, Liu Y Q and Xue Z Y. 2019a. Residual invertible spatio-temporal network for video super-resolution. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1): 5981-5988 [DOI: 10.1609/aaai.v33i01.33015981]

Zhu X Z, Hu H, Lin S and Dai J F. 2019b. 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

江俊君,男,教授,主要研究方向为计算机视觉、机器学习和图像处理。E-mail: jiangjunjun@hit.edu.cn

程豪,男,硕士研究生,主要研究方向为计算机视觉、机器学习和图像处理。E-mail: 1300314319@stu.hit.edu.cn

李震宇,男,硕士研究生,主要研究方向为计算机视觉、机器学习和图像处理。E-mail: 1170300110@stu.hit.edu.cn

刘贤明,男,教授,主要研究方向为计算机视觉、机器学习和图像处理。E-mail: csxm@hit.edu.cn

王中元,男,教授,主要研究方向为计算机视觉、机器学习和图像处理。E-mail: wzy_hope@163.com