基于自注意力机制的视频超分辨率重建

秦昊宇,葛 瑶,张力波,吴学致,任卫军
(长安大学 信息工程学院,陕西 西安 710064)

摘 要:现有的视频超分辨率重建方法虽然对提高视频分辨率取得了良好效果,但是很多方法没有充分考虑视频帧间运动时间域与空间域的关联性。针对这个问题,提出一种融合时间和空间域的视频超分辨率重建模型 VTSSR,用于在同一个网络模型中同时对视频进行时间和空间域超分辨率重建。该模型使用卷积层和多个残差块对低帧率、低分辨率视频进行特征提取,通过特征插值生成中间帧的特征图,采用改进的基于自注意力机制模块同时融合特征图时间和空间信息,采用亚像素卷积上采样重建得到高帧率的高分辨率视频。VTSSR 模型在 Vid4 数据集测试表明,其能够克服光流预测难以处理遮挡、复杂运动的局限性,还能解决不同相邻帧对于关键帧重建贡献不同的问题,提高了视频超分辨率重建水平。

关键词:视频超分辨率重建;深度学习;残差神经网络;视频插值;多对齐融合;自注意力机制
中图分类号:TP399 **文献标识码**:A **文章编号**:1673-629X(2022)08-0042-07
doi:10.3969/j.issn.1673-629X.2022.08.007

Video Super-resolution Reconstruction Based on Self Attention Mechanism

QIN Hao-yu,GE Yao,ZHANG Li-bo,WU Xue-zhi,REN Wei-jun
(School of Information Engineering,Chang'an University,Xi'an 710064,China)

Abstract:Although the existing video super-resolution reconstruction methods have achieved excellent results in improving video resolution,many methods do not fully take into account the correlation between video frame motion time domain and space domain. To solve this problem,a video super-resolution reconstruction model VTSSR integrating time and space domain is proposed to reconstruct video in time and space domain at the same time in the same network model. The model uses convolution layer and multiple residual blocks to extract the features of low frame rate and low resolution video,generates the feature map of intermediate frame through feature interpolation,uses the improved self attention mechanism module to fuse the temporal and spatial information of the feature map at the same time, and uses sub-pixel convolution up sampling to reconstruct the high frame rate and high resolution video. The test of VTSSR model on Vid4 data set shows that it can overcome the limitations of optical flow prediction that it is difficult to deal with occlusion and complex motion,solve the problem of different contributions of different adjacent frames to key frame reconstruction,and improve the level of video super-resolution reconstruction.

Key words:video super-resolution reconstruction;deep learning;residual neural network;video interpolation;multi alignment fusion;self attention mechanism

0 引 言

随着智能手机及各类摄影摄像设备的普及,图像、视频在人们生活中占据着越来越重要的地位。同时,人们对于图像、视频清晰度的需求也在逐渐提高。自 Harris 和 Goodman 首次提出图像超分辨率重建的概念与方法[1]以来,超分辨率方法作为计算机视觉领域中图像处理的一项技术,能够提高已经拍摄出的图像视频的分辨率,已经广泛应用到人们生活中的各个方面,具有重要的研究价值。

超分辨的核心在于寻找低分辨率图像与高分辨率图像特征之间的映射关系[2]。Dong Chao 等人在 2014 年首次将深度卷积网络 CNN 融入图像超分辨重建,提出了一种全卷积网络模型 SRCNN[3]。随着深度学习的不断深入研究及其在图像处理应用范围的扩大,更多基于深度学习方法的超分辨率重建网络模型正在进一步发挥作用[4],而超分辨率技术也不再局限于图像,而是开始向视频领域发展。视频超分辨率重建可分别在时间域和空间域上进行重建[5]。空间视频超分辨率

收稿日期:2021-09-10 **修回日期**:2022-01-11
基金项目:陕西省重点研发项目(2021GY-033)
作者简介:秦昊宇(1998-),女,在读硕士,研究方向为图像处理、图像超分辨率重建;任卫军,副教授,硕导,研究方向为图像视频处理。

重建通过引入更多的相邻帧,将帧间互补信息对齐融合到关键帧以提高关键帧重建效果。时间可变形的对齐网络(temporally deformable alignment network, TDAN)[6]对从原始帧提取的特征使用可变形卷积网络,自适应地完成当前帧与相邻帧的对齐,并动态地根据估计出的特征空间补偿信息进行隐式运动补偿,从而通过重建模块得到高分辨率的视频帧。EDVR[7](enhanced deformable convolutional networks)在 TDAN 模型的基础上提出了多尺度特征图对齐模块,更好地完成了帧间互补信息对齐。时间域的视频超分辨率重建主要是通过给定当前帧图像和下一帧图像,从而生成中间帧的视频插帧技术来实现。目前,通过深度神经网络学习现有帧与未知帧的映射关系存在一定困难,因此常用通过学习得到的中间帧的光流信息来进行传统插值,进而生成中间帧。该文是在现有的算法研究基础上对视频超分辨率重建进行深入研究,构建了一种融合时间与空间域的视频超分辨率重建模型 VTSSR,实验证明,该模型充分考虑到了视频帧间运动时间与空间的关联性,提高了视频超分辨率的重建效果。

1　超分辨率重建理论

现实生活中,各种外在影响会使得采样得到低分辨率图像,这一现象称为图像退化[8]。逆向处理图像退化,从而恢复出高分辨率图像和视频的技术就称为超分辨率重建技术。

视频超分辨率重建分为时间域和空间域的视频超分辨率重建。时间域超分辨率重建主要是将由于采样设备、视频压缩等造成高频信息丢失的低帧率视频帧重构成高帧率[9],重构过程主要使用视频插帧的方法。常用的视频插帧方法主要是基于运动补偿的视频插帧,其主要思想是通过运动估计和运动补偿在原视频序列连续的两帧之间插入图像帧,基于运动补偿的插帧方法步骤如图 1 所示。

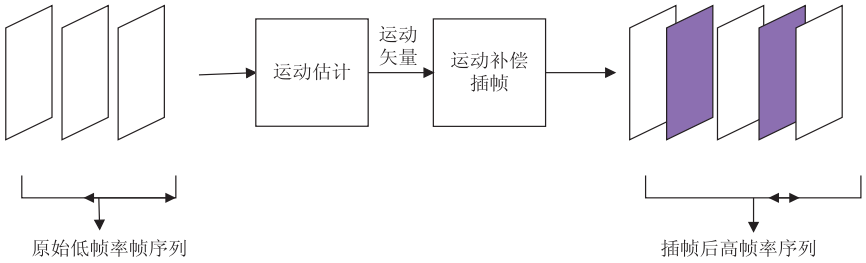

图 1　基于运动补偿插帧

如图 1 所示,该过程将输入的原始低帧率视频序列经过运动估计得到相邻帧之间的运动矢量,再将这些运动矢量经过运动补偿插帧等操作生成插帧后的高帧率序列。其中,运动估计是计算视频帧上同一个像素点在相邻图像帧之间运动时发生的空间偏移量[10],而运动补偿就是根据这些偏移量计算该像素点在中间帧的对应位置,从而补偿差值生成中间帧。

空间域超分辨率重建是从连续的低分辨视频帧序列中重建得到对应的高分辨率帧序列,其中重构过程使用多帧图像超分辨率重建方法实现,其关键在于多帧帧间信息的配准。这些由视频得到的多帧图像在亮度和像素上存在细微差别,能够通过捕获帧间差异信息完成超分辨率重建[11],方法过程如图 2 所示。

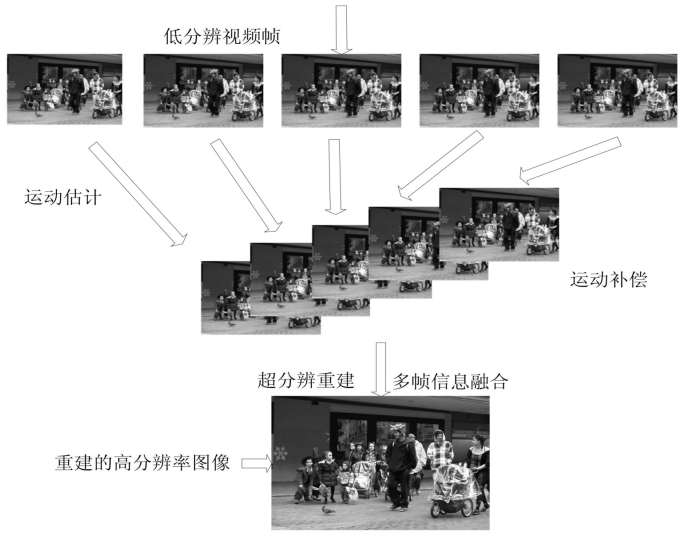

图 2　视频超分辨率重建

由图 2 可知,该过程首先将输入的原始低分辨率视频帧序列利用运动估计算法预测帧间运动矢量,再对本段序列帧中的关键帧进行运动补偿,接着将相邻帧图像与当前关键帧图像进行对齐配准,使两帧位于同一坐标系中,最后经过重建网络将多个特征图像融合得到关键帧的高分辨率图像。

2 融合时间域与空间域的视频超分辨率重建模型

2.1 模型框架

融合时间域和空间域的视频超分辨率重建模型中的重建网络由特征提取、特征插值、自注意力机制融合以及亚像素卷积上采样和残差块四部分组成,能够同时对视频进行时间域和空间域超分辨率重建,最后输出高帧率的高分辨率视频。由于视频帧间互补信息在时间和空间上具有一定的关联性,该模型采用可变形卷积对齐、自注意力融合技术增强了这一关联性,从而进一步提升了视频超分辨率重建的效果。模型框架如图 3 所示。

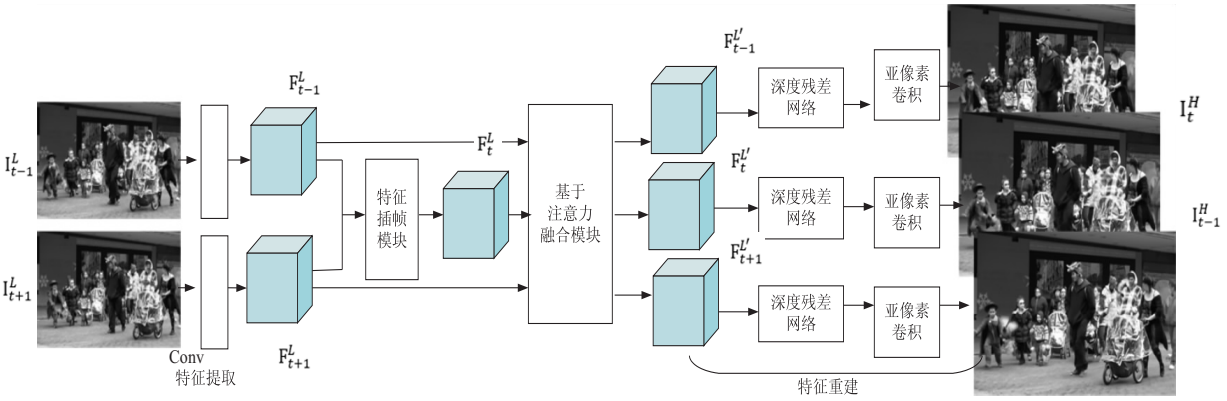

图 3 VTSSR 模型结构

2.2 特征提取

基于深度学习的视频超分辨率重建使用卷积神经网络,区别于光流法复杂的假设条件和公式推理,直接通过卷积运算来学习相邻帧之间的运动信息,降低了特征提取的复杂度,提高了特征的语义性。此外,引入残差块能够加深卷积结构,从而增强模型重建能力并提高重建视频质量。该文使用卷积神经网络进行特征提取的模块如图 4 所示。

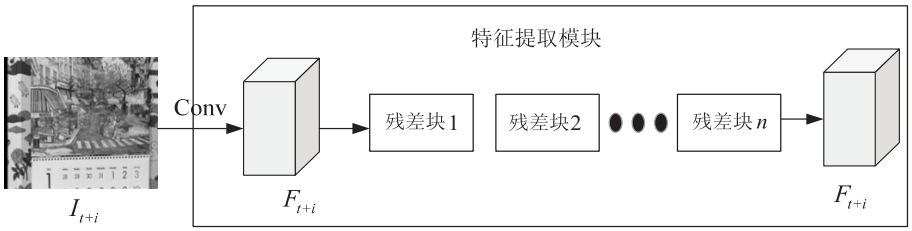

图 4 特征提取模块

图 4 中展示了该特征提取模块的操作流程,该模块由一个卷积层和多个残差块组成。首先,输入的相邻奇数帧 I_{t+i} 经过卷积网络提取出其特征图 F_{t+i},再将提取的特征图 F_{t+i} 经过残差块提高模型重建能力,最后将其用于特征插帧模块的输入。

2.3 特征插值

由于该模型是融合了时间域和空间域的视频超分辨率重建模型,模型关键在于增强视频帧间互补信息在时间和空间上的关联性,因此特征时间插值模块在相邻帧的特征图上直接对齐得到中间帧特征图,而不是先重建生成中间帧再得到中间帧的特征图。TDAN[12] 基于可变形卷积的对齐模块,利用卷积网络学习帧间运动进行信息建模代替了光流预测法,能够更好地对齐相邻特征图。融合了时间域和空间域的视频超分辨率重建模型参考了 TDAN 提出的原理,采用同模型不同方向的卷积,并加入了特征时间插值模块的设计,如图 5 所示。

图 5 中,模块输入为相邻的两帧特征图 F_{t+1} 和 F_{t-1},待插中间帧的特征图为 F_t,通过学习一个特征时间插值函数 $f(\cdot)$ 并利用相邻的两帧特征图 F_{t+1} 和 F_{t-1} 直接合成和待插中间帧的特征图 F_t,它们之间的关系用公式表示为:

$$F_t = f(F_{t-1}, F_{t+1}) = H(F_{t-1\to t}, F_{t+1\to t}), t \in 2N \quad (1)$$

其中,$F_{t+1\to t}$ 为 F_{t+1} 到中间帧特征图 F_t 的对齐特征,$F_{t-1\to t}$ 为 F_{t-1} 到中间帧特征图 F_t 的对齐特征,$H(\cdot)$ 为一聚合采样特征的混合函数。

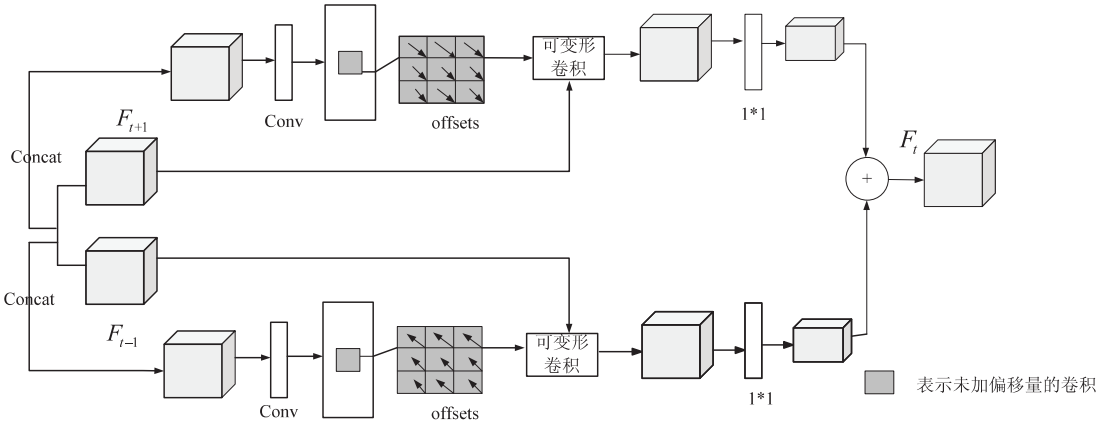

图 5　特征时间插值模块

由于中间帧特征图 F_t 还未合成,无法直接计算 F_t 与 F_{t-1} 、F_t 与 F_{t+1} 之间的运动信息,因此使用可变形采样函数来隐式地获取 F_{t-1} 与 F_{t+1} 之间的运动信息,以此来近似代替。首先输入 F_{t+1} 和 F_{t-1} 合并 Concat 后经过 3 * 3 的卷积层,目的是减少通道数降低参数量。然后通过一个卷积层去预测输出通道数量为 | R | 的采样参数 θ_1,如式(2)所示:

$$\theta_1 = g_1(F_{t-1}, F_{t+1}) \tag{2}$$

其中,采样参数 θ_1 为可偏移学习量,g_1 表示多个卷积层的一般函数。同理,输入 F_{t-1} 和 F_{t+1} ,Concat 后经过 3 * 3 的卷积层,然后通过一个卷积层去预测输出通道数量为 | R | 的采样参数 θ_2,如式(3)所示:

$$\theta_2 = g_2(F_{t-1}, F_{t+1}) \tag{3}$$

接着使用可变形卷积计算出 F_{t+1} 到中间帧特征图 F_t 的对齐特征 $F_{t+1\to t}$ 和 F_{t-1} 到中间帧特征图 F_t 的对齐特征 $F_{t-1\to t}$,如式(4) 所示:

$$\begin{aligned} F_{t+1\to t} &= T(F_{t+1}, \theta_1) = \mathrm{DConv}(F_{t+1}, \theta_1) \\ F_{t-1\to t} &= T(F_{t-1}, \theta_1) = \mathrm{DConv}(F_{t-1}, \theta_2) \end{aligned} \tag{4}$$

最后,得到的两个对齐特征通过 $H(\cdot)$ 混合函数分别相乘 1 * 1 卷积层再对位相加得到了最终中间帧的特征图 F_t 。

2.4　特征融合

该模型在特征融合模块中引入自注意力机制,将通过联合学习对齐得到的多特征图的时间、空间维度特征信息赋予新的权重并重新分配,能够自适应地学习特征信息在不同维度之间的关联性。基于这种方法的特征融合,能够增强有用特征而抑制无用特征,从而更好地处理特征的帧间运动并利用帧内信息在时间和空间上挖掘特征信息。其结构如图 6 所示。

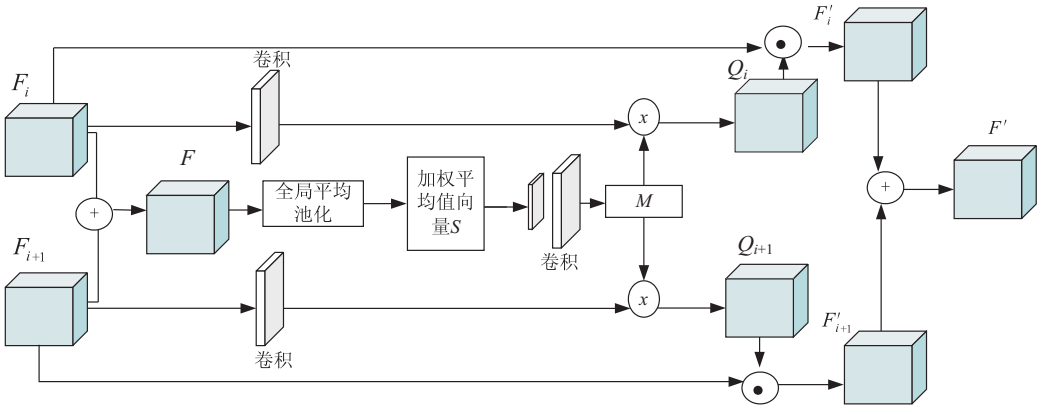

图 6　基于自注意力融合模块

如图 6 所示,首先将需要融合的特征图序列 F_i 合并 concat 后得到全部特征图的和 F ,然后将特征图通过全局池化层得到全部通道数的加权平均值向量 $\boldsymbol{S}$,其中 $\boldsymbol{S}$ 的通道数为 C 个,计算公式如式(5)所示:

$$S_c = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} F_c(w, h) \tag{5}$$

接着使用两个全连接层学习全部通道间的相关性,前一个全连接层使用压缩因子 r 压缩通道数 C/r ,后一个全连接层将通道数扩为原始的通道数 C ,计算得到 F 的特征向量矩阵 $\boldsymbol{M}$,计算公式如式(6)所示:

$$\boldsymbol{M} = \boldsymbol{W}_2(\sigma(\boldsymbol{W}_1 \cdot F)) \tag{6}$$

其中, $\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 是两个全连接层的权重矩阵。这样先压缩后扩充通道的方法能够给各通道分配各自的注意力,从而可以达到增强有用特征而抑制无用特征的目

的。接着使用卷积核为 1 * 1 的卷积层来卷积压缩特征图的时间维,将 F_i 的尺寸从 $C*H*W$ 变为 $H*W$,学习每个输入特征矩阵在空间维度上的内部相关性得到序列$\{q_i\}$计算公式如式(7)所示:

$$q_i = \mathrm{CNN}(\boldsymbol{W}_3 \cdot F_i) \tag{7}$$

式中,$\boldsymbol{W}_3$ 是卷积层的权重矩阵。

将得到的$\{q_i\}$和特征向量矩阵 $\boldsymbol{M}$ 进行向量点乘得到通道和空间的相关性$\{P_i\}$计算公式,如式(8)所示:

$$P_i = q_i \cdot \boldsymbol{M} \tag{8}$$

使用激活函数 sigmod 得到更加突出重要元素的权重矩阵$\{g_i\}$,如式(9)所示:

$$g_{i,w,h,c} = \frac{\mathrm{e}^{P_{i,w,h,c}}}{\sum_{k}^{n} \mathrm{e}^{P_{k,w,h,c}}} \tag{9}$$

式中,(w,h,c)指像素的空间坐标和通道位置。最后将权重矩阵和输入特征 F_i 对位求和得到最终的融合特征图:

$$\hat{F} = \sum_{i=1}^{n} b_i \odot U_i \tag{10}$$

2.5 重建模块

现有常见的基于深度学习的重建网络主要是提取出低分辨率图像中的特征,并通过学习这些特征将它们扩大成高分辨率图像,从而实现超分辨率重建。该文提出的融合时间域与空间域的视频超分辨率重建模型在重建时使用高分辨率图像 HR 重建模块,该模块输入来自低分辨率图像中提取深层特征后构成的特征图序列,使用多个堆叠残差块和两个卷积网络将输入的低分辨、高帧率的特征图像序列转为高分辨、高帧率图像输出。

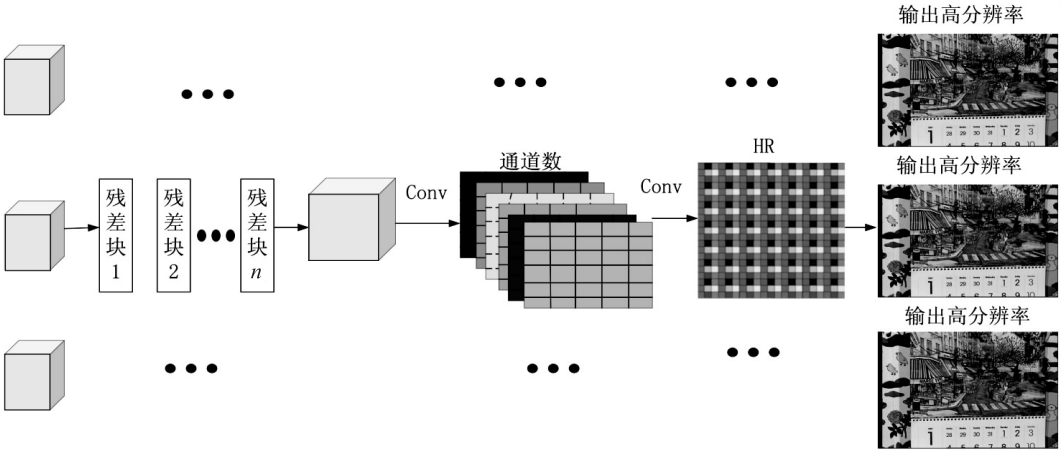

图 7 HR 重建模块

由图 7 可知,HR 重建模块对输入的特征图序列使用残差块组成的网络来继续学习特征,这种方法能够更好地利用图像高频信息[13],减小运算量。最后利用亚像素卷积上采样扩大特征图尺寸,输出重建的目标分辨率图像。

3 实验结果与分析

3.1 实验数据集

为了验证该文提出的融合时间域与空间域的视频超分辨重建模型的有效性,使用视频超分辨领域主流的数据集 Vimeo-90k[14]、Vid4[15] 对模型进行实验验证,并将实验结果和目前的视频超分辨算法模型进行比较。本模型在获取低分辨率图像数据集时采用双三次插值下采样方法,测试时对放大 4 倍的图像进行模型重建,并使用主观视觉和客观指标峰值信噪比 PSNR、结构相似性 SSIM[16] 对模型重建效果进行评价。

3.2 实验处理及参数设置

由于下载的公开数据集主要是高分辨率视频,而训练时需要与之对应的低分辨率视频,因此使用 Matlab 工具将高分辨率视频下采样缩小成与之对应的低分辨率图像,再分别将这些高分辨率图像和低分辨率图像生成 lmdb 格式以便作为输入数据。本次实验使用视频序列中奇数标签作为输入,训练模型配置文件参数如表 1 所示。

表 1 配置文件参数设置

参数	数值
训练批次	16
下采样因子	4
低分辨率输入图像尺寸	32 * 32
GPU 个数	1
日志书写频率	2 000/次
初始学习率	4×10^{-4}

模型中使用的损失函数如式(11)所示:

$$L_{\mathrm{rec}} = \sqrt{\| I_t^{\mathrm{GT}} - I_t^{H} \|^2 + \varepsilon^2} \tag{11}$$

式中,L_{rec} 为每个样本对的总损失,I_t^{GT} 为输出的图像,I_t^{H} 为训练样本中对应的高分辨率图像。常量 ε 取值 1e-3 能够确保训练过程中数据稳定。

3.3　实验结果分析

训练完成后,在 Vid4 和 Vimeo-90k 数据集上对构建的融合时间域和空间域视频超分辨率重建模型 VTSSR 的可行性和有效性进行验证。

3.3.1　PSNR 和 SSIM 指标值

从 Vid4 数据集和 Vimeo-90k 训练集中抽取测试集,对超分辨倍数为 4 的样本进行实验,与两种时间插帧模型 Sepconv[17]、DAIN[18] 和三种空间超分辨模型 Biubic[19]、RCAN[19]、EDVR[6] 联合的方法进行对比。VTSSR 是该文构建的融合时间域和空间域的视频超分辨率重建模型。通过对比可以看出,该模型在一定程度上优于其他模型,各自模型的峰值信噪比 PSNR 和结构相似性 SSIM 指标对比如表 2 所示。

表 2　模型客观指标对比(PSNR(DB)/SSIM)

模　型		Vid4		Vimeo-90k	
VFI	VSR	PSNR	SSIM	PSNR	SSIM
Sepconv	Biubic	23.51	0.627 3	30.61	0.863 3
Sepconv	RCAN	24.92	0.723 6	33.59	0.912 5
Sepconv	EDVR	25.93	0.779 2	34.22	0.924 0
DAIN	Biubic	23.55	0.626 8	30.67	0.863 6
DAIN	RCAN	25.03	0.726 1	33.82	0.914 6
DAIN	EDVR	26.12	0.778 4	34.66	0.928 1
VTSSR		26.29	0.795 6	35.79	0.937 4

3.3.2　主观对比

为了进一步对比重建效果,在 Vimeo-90ktest 数据集选取多个不同的视频片段的图像,对这些图像进行模型重建,并在主观视觉上对高分辨率图像、下采样低分辨率图像、SepConv + EDVR、DAIN + Bicubic、DAIN + RCAN 与构建的模型 VTSSR 的重建结果进行评价对比;在 Vid4 数据集上将高分辨率图像、下采样低分辨率图像与构建的模型 VTSSR 的重建结果进行评价对比。对比结果如图 8 所示。

图 8　Vimeo-90ktest 和 Vid4 测试集主观视觉对比

图 8 中,选取了 Vimeo-90ktest 和 Vid4 数据集中各一个视频片段图像,通过对比可以看出 VTSSR 模型主观视觉优于其他模型重建的效果。

通过比较不同重建模型在相同测试集上的结果,提出的融合时间域和空间域的视频超分辨率重建模型 VTSSR 在量化指标和观察主观效果上都有一定的

优势。

4 结束语

为提高视频分辨率,构建了一种融合时间域和空间域的视频超分辨率重建模型 VTSSR,可以在同一个网络模型中同时对视频进行时间域和空间域超分辨率重建。该模型以低帧率的低分辨率视频作为输入,首先,使用卷积层和多个残差块进行特征提取,使用帧特征插值生成中间帧的特征图;其次,采用改进的基于自注意力机制模块,融合特征图时间和空间信息;最后,采用亚像素卷积上采样重建,输出高帧率的高分辨率视频。在 Vimeo-90ktest 和 Vid4 测试集上的测试表明,该模型能够克服光流预测难以处理遮挡、复杂运动的局限性、解决不同相邻帧对于关键帧重建贡献不同的问题。在 Vimeo-90ktest 测试集上其峰值信噪比为 35.79 dB,结构相似性为 0.937 4;在 Vid4 测试集上其峰值信噪比为 26.29 dB,结构相似性为 0.795 6,与其他重建模型相比均有提高。

参考文献:

[1] 谢 旺. 基于学习的视频超分辨率重建研究[D]. 广州:华南理工大学,2019.

[2] 曹 斌. 基于深度学习的超分辨重建[D]. 西安:西安电子科技大学,2017.

[3] 唐艳秋,潘 泓,朱亚平,等. 图像超分辨率重建研究综述[J]. 电子学报,2020,48(7):1407-1420.

[4] HE X,YANG L,TENG Q,et al. Learning-based super-dimension (SD) reconstruction of porous media from a single two-dimensional image[C]//2016 IEEE international conference on signal processing, communications and computing (ICSPCC). Hong Kong,China:IEEE,2016:1-5.

[5] 杜珊珊. 基于视频帧图像的超分辨率重建应用算法研究[D]. 济南:山东师范大学,2018.

[6] TIAN Y,ZHANG Y,FU Y,et al. TDAN:temporally-deformable alignment network for video super-resolution[C]//2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR). Seattle,WA,USA:IEEE,2020:3357-3366.

[7] WANG X, CHAN K, YU K,et al. EDVR:video restoration with enhanced deformable convolutional networks[C]//2019 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW). Long Beach,CA,USA:IEEE,2019:1954-1963.

[8] KATSAGGELOS A K,MOLINA R,MATEOS J. Super resolution of images and video[J]. Synthesis Lectures on Image Video & Mulitmedia Processing,2012,3(1):682-689.

[9] 张震洲,高 昆,李 维,等. 光学遥感图像的超分辨率处理技术综述[J]. 航天返回与遥感,2020,41(6):21-33.

[10] 李定一. 基于深度学习的视频超分辨率算法研究[D]. 合肥:中国科学技术大学,2019.

[11] LEE C M,YEH H F. Adaptive band-based super-resolution reconstruction of video sequence[C]//IEEE international conference on applied system innovation. Sapporo, Japan:IEEE,2017:288-291.

[12] TAO X,GAO H,LIAO R,et al. Detail-revealing deep video super-resolution[C]//2017 IEEE international conference on computer vision (ICCV). Venice, Italy:IEEE,2017:4482-4490.

[13] 杜 鹏. 面向视频超分辨率的深度学习研究[D]. 成都:电子科技大学,2020.

[14] XUE T,CHEN B,WU J,et al. Video enhancement with task-oriented flow[J]. International Journal of Computer Vision,2019,127(8):1106-1125.

[15] LIU C,SUN D. On Bayesian adaptive video super resolution[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2014,36(2):346-360.

[16] FISCHER P,DOSOVITSKIY A,ILG E,et al. FlowNet:learning optical flow with convolutional networks[C]//2015 IEEE international conference on computer vision (ICCV). Santiago, Chile:IEEE,2016:2758-2766.

[17] GAO Y,KOCH R,BREGOVIC R,et al. IEST:interpolation-enhanced shearlet transform for light field reconstruction using adaptive separable convolution[C]//2019 27th European signal processing conference (EUSIPCO). A Coruna, Spain:IEEE,2019:1-5.

[18] BAO W,LAI W S,MA C,et al. Depth-aware video frame interpolation[C]//IEEE/CVF conference on computer vision & pattern recognition. Long Beach,CA,USA:IEEE,2019:3698-3707.

[19] KEYS R G. Cubic convolution interpolation for digital image processing[J]. IEEE Transactions on Acoustics Speech and Signal Processing,2003,29(6):1153-1160.