Journal of Computer Applications
计算机应用,2020,40(3):819－824
ISSN 1001-9081
CODEN JYIIDU
2020-03-10
http://www.joca.cn

文章编号:1001-9081(2020)03-0819-06
DOI:10.11772/j.issn.1001-9081.2019081474

基于深度体素流的模糊视频插帧方法

林传健[1,2],邓　炜[3],童　同[3],高钦泉[1,2,3*]
(1.福州大学 物理与信息工程学院,福州 350116;　2.福建省医疗器械与医药技术重点实验室(福州大学),福州 350116;
3.福建帝视信息科技有限公司,福州 350116)
(*通信作者电子邮箱gqinquan@fzu.edu.cn)

摘　要:针对视频运动模糊严重影响插帧效果的情况,提出了一种新型的模糊视频插帧方法。首先,提出一种多任务融合卷积神经网络,该网络结构由两个模块组成:去模糊模块和插帧模块。其中,去模糊模块采用残差块堆叠的深度卷积神经网络(CNN),提取并学习深度模糊特征以实现两帧输入图像的运动模糊去除;插帧模块用于估计帧间的体素流,所得体素流将用于指导像素进行三线性插值以合成中间帧。其次,制作了大型模糊视频仿真数据集,并提出一种先分后合、由粗略至细致的训练策略,实验结果表明该策略促进了多任务网络有效收敛。最后,对比前沿的去模糊和插帧算法组合,实验指标显示所提方法合成中间帧时峰值信噪比最少提高1.41 dB,结构相似性提升0.020,插值误差降低1.99。视觉对比及重制序列展示表明,所提模型对于模糊视频有着显著的帧率上转换效果,即能够将两帧模糊视频帧端对端重制为清晰且视觉连贯的三帧视频帧。

关键词:深度学习;卷积神经网络;去模糊;体素流;视频插帧
中图分类号:TP391;TP183　　**文献标志码**:A

Blurred video frame interpolation method based on deep voxel flow

LIN Chuanjian[1,2], DENG Wei[3], TONG Tong[3], GAO Qinquan[1,2,3*]
(1. *College of Physics and Information Engineering, Fuzhou University, Fuzhou Fujian* 350116, *China*;
2. *Key Lab of Medical Instrumentation and Pharmaceutical Technology of Fujian Province* (*Fuzhou University*), *Fuzhou Fujian* 350116, *China*;
3. *Imperial Vision Technology Company Limited, Fuzhou Fujian* 350116, *China*)

Abstract: Motion blur has an extremely negative effect on video frame interpolation. In order to handle this problem, a novel blurred video frame interpolation method was proposed. Firstly, a multi-task fusion convolutional neural network was proposed, which consists of a deblurring module and a frame interpolation module. In the deblurring module, based on the deep Convolutional Neural Network (CNN) with stack of ResBlocks, motion blur removal of two input frames was implemented by extracting and learning the deep blur features. And the frame interpolation module was used to estimate voxel flow between two consecutive frames after blur removal, then the obtained voxel flow was used to guide the trilinear interpolation of the pixels to synthesize the intermediate frame. Secondly, a large blurred video simulation dataset was made, and a "first separate and then combine" "from coarse to fine" training strategy was proposed, experimental results show that this strategy promotes the effective convergence of the multi-task fusion network. Finally, compared with the simple combination of the state-of-the-art deblurring and frame interpolation algorithms, experimental metrics show that the intermediate frame synthesized by the proposed method has the peak-to-noise ratio increased by 1.41 dB, the structural similarity improved by 0.020, and the interpolation error decreased by 1.99, at least. Visual comparison and reconstructed sequences show that the proposed model performs good frame rate up conversion effect for blurred videos, in other words, two blurred consecutive frames can be reconstructed end-to-end to three sharp and visually smooth frames by the model.

Key words: deep learning; Convolutional Neural Network (CNN); debluring; voxel flow; video frame interpolation

0　引言

随着科技发展,生活水平提升,人们可以随时随地观看视频,对视频质量的要求也越来越高。而视频插帧正是计算机视觉在视频增强中的一个典型应用。视频是由一系列内容连续的静态图像按照一定的时序连贯组成,插帧任务将视频中每每两帧连续帧作为输入,利用帧间信息预测出中间帧,从而重制出帧率更高且画面更加连贯的视频。对于视频插帧这个研究热点,难点在于如何快速精确生成中间帧,使源序列在视觉上更加平滑过渡。

近年来随着人工智能浪潮的兴起,各种基于深度学习的视频插帧思路频频提出。FlowNet[1-2]方法最早证明了用卷积神经网络直接预测帧间光流的可行性,但是在一些特定情况下估计所得光流的精度不够理想(如运动过大、目标遮挡、光

收稿日期:2019-08-29;**修回日期**:2019-10-23;**录用日期**:2019-10-23。　**基金项目**:国家自然科学基金资助项目(61802065)。
作者简介:林传健(1994—),男,福建宁德人,硕士研究生,主要研究方向:计算机视觉、图像处理;　邓炜(1992—),男,福建龙岩人,硕士,主要研究方向:计算机视觉、图像处理;　童同(1986—),男,安徽安庆人,研究员,博士,主要研究方向:人工智能与计算机视觉、医学图像处理与分析;　高钦泉(1986—),男,福建福清人,副研究员,博士,主要研究方向:人工智能与计算机视觉、医学图像处理与分析、计算机辅助手术导航。

线剧烈变化、模糊等),因此基于光流的插帧方法常常出现像素混乱。另外缺乏准确的光流真值用于训练也是光流法的一个弊端。Niklaus 等[3-4]通过估计能够用于捕获帧间运动与插值系数的空间自适应卷积核以合成中间帧,该方法无需光流监督,且能够较好地应对局部的运动变化。Liu 等[5]提出用一个卷积神经网络估计出帧间紧密的体素流,基于体素流直接从两帧输入参考帧中拷贝相同的像素,使得合成帧更加清晰真实,同时大大节省了计算量。Jiang 等[6]对帧间的双向光流进行线性融合从而能够合成任意时间节点的中间帧,在视频慢动作应用上效果亮眼。另有研究者探究如何融合利用图像的深度、背景、边缘等[7-9]信息更好地保护图像结构,以获得更加清晰的插帧效果。上述这些方法虽然能够初步良好估计出帧间物体的小幅运动,对于画中目标遮挡情形的处理也有了不小的提升,但针对视频画面模糊的情况进行插帧则研究甚少,而往往视频中的运动模糊会大大影响插帧效果,甚至引起中间生成帧的画面全盘崩坏。

图像盲去模糊是一种高度病态问题,许多先驱工作从各种角度出发尝试解决[10]。尤其近年来深度学习的发展,同样带动图像去模糊任务取得不少新的突破。Kupyn 等[11]将生成对抗网络的思想应用到去模糊任务之中,其图像恢复细节更加丰富生动;Nah 等[12]利用多尺度信息融合以学习图像模糊特征,效果显著但是网络相对复杂;Tao 等[13]在此基础上通过长短时记忆(Long Short-Term Memory, LSTM)级联多尺度间的模糊-清晰对加强特征学习。图像去模糊技术开始突破特定场景壁垒慢慢走向实用。

综上所述,深度学习带来的多种计算机视觉任务突破为多任务融合的端对端方法打下坚实的基础,本文即在此背景下,提出针对模糊输入帧的高质量插帧重制方法,设计了一个完整有序的深度卷积神经网络,其中去模糊网络模块用于提升源图像质量,插帧网络模块对去除模糊之后的参考帧提取有效的帧间信息,依据可靠的体素流估计出理想的中间帧。

本文的主要内容如下:1)针对模糊视频影响插帧任务性能的背景,制作了适用于模糊插帧的训练与测试数据集,提出了一种新型的去模糊与插帧相结合的网络结构;2)训练并结合了图像去模糊模型与视频插帧模型。设计了合适的损失函数和分步联合训练策略,使模型有效收敛;3)对比两组前沿的去模糊算法与插帧算法组合以及本文所提出联合训练方法两个阶段的模型,客观评估指标的提升与可靠的视觉结果验证本文的工作能够将两帧连续模糊输入帧有效地端对端重制为三帧连贯且清晰的视频帧序列。

1 网络设计方案

借鉴最新的研究成果,综合考虑性能,设计了新型的网络结构。整体网络主要由去模糊模块与插帧模块两部分组成。首先将内容连续的两帧输入帧分别通过去模糊网络生成对应清晰帧,接着通过插帧网络估计出清晰帧之间的体素流,再经过三线性插值计算得到中间帧。下面将详细阐述整个网络的结构组成。

1.1 去模糊模块

关于运动模糊的去除,已有不少可供参考的研究,在去模糊任务中,需要获取足够大的感受野以应对剧烈的运动,而网络一味加深又将造成特征图过小而难以恢复,参数数量急剧上升导致收敛速度下降等负面效应。U-Net[14]网络结构最早应用于医学图像分割任务,编码器与解码器对称堆叠组合的特殊设计如今也被广泛应用于视频超分辨率、图像修复等其他计算机视觉任务[15-17]。残差网络和跳跃连接由 He 等[18]提出,通过残差学习可以使网络中各层的梯度信息尽可能被传递,缓解深层网络中的梯度消失问题。经过实验测试权衡,本文的去模糊模块采用了一种适量残差块(ResBlock)堆叠的 U-Net 变形网络结构。残差块的基本结构如图 1 所示。

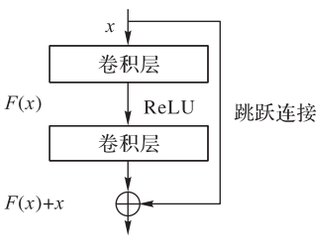

图 1 残差块示意图

Fig. 1 Schematic diagram of ResBlock

如图 1 所示,每个残差块包含两个卷积核数相同的卷积层,两个卷积层间经过 ReLU 非线性激活函数。该残差块可用式(1)表示:

$$H(x)=F(x)+x \tag{1}$$

其中:x 表示恒等映射,$F(x)$ 表示网络前向传播学习到的局部残差,通过跳跃连接将两者相加得到残差块的输出结果 $H(x)$,并作为下一个残差块的输入。残差块的适当堆叠能够有效提取模糊图像的深层特征,并且可以解决训练深层网络时的梯度消失问题,从而学习清晰图像与模糊图像之间的映射关系,最终实现高质量的图像去模糊功能,保证后续插帧处理初始的可靠性。去模糊模块框架如图 2 所示。

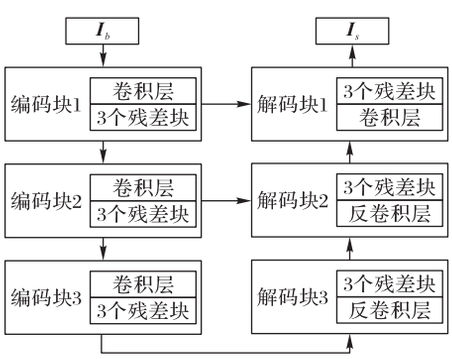

图 2 去模糊模块框架

Fig. 2 Debluring module framework

图 2 中 $\boldsymbol{I}_b$ 表示输入的模糊图像,$\boldsymbol{I}_s$ 表示输出的去模糊图像。该网络由编码块与解码块对称堆叠组合,编码块中的卷积层与对应解码块的反卷积层通过跳跃连接以融合不同尺度间的信息。每一个编码块都由一个卷积层和三个残差块依次排列组成,对应的解码块由三个残差块和一个反卷积层依次排列组成。其中每个卷积层与反卷积层后都接有非线性激活函数 ReLU,且卷积核大小均为 5×5,以便通过更大的感受野获得更丰富的全局特征,同时步长设置为 1,以保证输出特征图与原图大小一致。具体的各卷积层参数设置如表 1 所示。整个过程可用式(2)、(3)近似表示:

$$\boldsymbol{f}=N_E(\boldsymbol{I}_b;\ \boldsymbol{\theta}_E) \tag{2}$$

$$\boldsymbol{I}_s=N_D(\boldsymbol{f};\ \boldsymbol{\theta}_D) \tag{3}$$

其中:$N_E(\cdot)$ 和 $N_D(\cdot)$ 分别表示带有超参数 $\boldsymbol{\theta}_E$ 与 $\boldsymbol{\theta}_D$ 的编码器部分网络与解码器部分网络,$\boldsymbol{f}$ 表示模糊图像 $\boldsymbol{I}_b$ 经过带有三个编码块的编码器部分网络得到的特征图,$\boldsymbol{I}_s$ 即为特征图 $\boldsymbol{f}$ 经过

三个解码块输出的去模糊图像。后续在整体网络框架中(见图3)对去模糊模块框架作简单表示。

表1　去模糊网络各卷积层参数设置

Tab. 1　Parameter setting of each convolution layer in deblurring network

编/解码块	卷积核大小	通道数	激活函数
编码块1	5×5	32	ReLU
编码块2	5×5	64	ReLU
编码块3	5×5	128	ReLU
解码块3	5×5	128	ReLU
解码块2	5×5	64	ReLU
解码块1	5×5	32	ReLU

1.2　基于深度体素流的插帧模块

在插帧模块,将对两帧去除模糊之后的图像进一步做插帧处理。目前已有不少视频插帧算法可供参考,比如AdaConv和SepConv等[3-4]在引言中已作介绍。深度体素流(Deep Voxel Flow, DVF)模型[5]通过卷积神经网络从两帧连续视频帧预测出体素流图和掩膜,继而引导图像之间的各像素合成更加真实的中间帧,本文尝试借鉴该方法的思想指导本文插帧部分。

整体网络框架如图3所示,将经过去除模糊的两帧连续帧联结作为输入,插帧模块部分的网络结构共有三对编-解码块和一个瓶颈块,每一对编-解码块经过跳跃连接加强信息传递,除了第一个编码块单独由卷积层构成,每个编码块和瓶颈块由卷积层与最大池化层组成,解码块由反卷积层与卷积层组成,各卷积层的参数设置如表2所示。最后通过一个卷积核大小为1×1的卷积层估计得到通道数为三的帧间体素流。

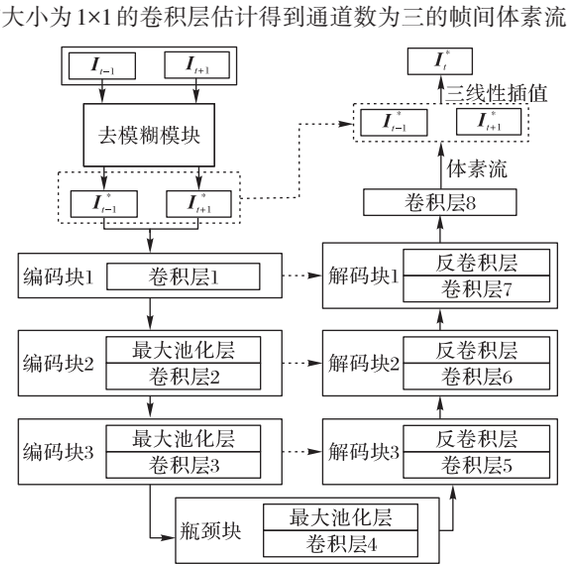

图3　整体网络框架

Fig. 3　Overall network framework

体素流$\boldsymbol{F}$的空域成分代表相邻帧间的光流,时域成分代表两个参考帧对于合成中间帧像素的线性混合权重。依据体素流$\boldsymbol{F}$对相邻两帧进行三线性插值合成中间帧,此过程可近似表示如下:

$$\boldsymbol{F}=Net(\boldsymbol{I}_{t-1}^*,\boldsymbol{I}_{t+1}^*;\boldsymbol{\omega}) \tag{4}$$

$$\boldsymbol{I}_t^*=\mathcal{T}_{x,y,z}(\boldsymbol{I}_{t-1}^*,\boldsymbol{I}_{t+1}^*,\boldsymbol{F}) \tag{5}$$

其中:$\boldsymbol{I}_{t-1}^*,\boldsymbol{I}_{t+1}^*$分别表示输入的两帧去模糊视频帧,$\boldsymbol{\omega}$为插帧网络$Net(\cdot)$中待学习的参数,$\boldsymbol{F}$表示网络输出的帧间体素流。$\mathcal{T}_{x,y,z}(\cdot)$为体素采样函数,$\boldsymbol{I}_t^*$表示最终的合成中间帧。体素通过三线性插值进行颜色重采样以合成中间估计帧,其原理示意图如图4所示。

表2　插帧网络各卷积层参数设置

Tab. 2　Parameter setting of each convolution layer in frame interpolation network

编/解码块	卷积核大小	通道数	激活函数
编码块1	5×5	64	ReLU
编码块2	5×5	128	ReLU
编码块3	3×3	256	ReLU
瓶颈块	3×3	256	ReLU
解码块3	3×3	256	ReLU
解码块2	5×5	128	ReLU
解码块1	5×5	64	ReLU
卷积层8	1×1	3	tanh

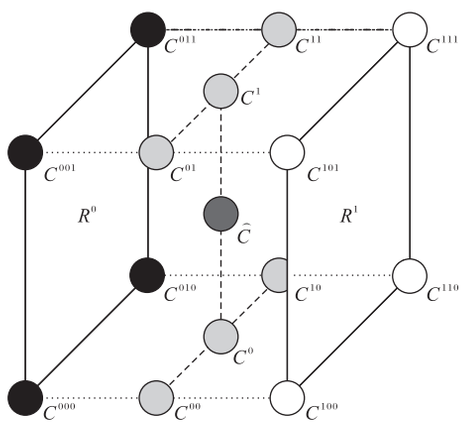

图4　三线性插值示意图

Fig. 4　Schematic diagram of trilinear interpolation

下面阐述体素采样的过程。如图4所示,给定一个体素,前后参考帧光流映射的对应区域有$(\boldsymbol{R}^0,\boldsymbol{R}^1)\in(\boldsymbol{I}_{t-1}^*,\boldsymbol{I}_{t+1}^*)$,坐标区间可分别表示为式(6):

$$\begin{cases}\boldsymbol{R}^0=(x-\Delta x,y-\Delta y)\\ \boldsymbol{R}^1=(x+\Delta x,y+\Delta y)\end{cases} \tag{6}$$

每个区域的4个顶点C^{ijk}分别表示如式(7):

$$\begin{cases}C^{000}=(\lfloor\boldsymbol{R}_x^0\rfloor,\lfloor\boldsymbol{R}_y^0\rfloor,0)\\ C^{001}=(\lfloor\boldsymbol{R}_x^0\rfloor,\lceil\boldsymbol{R}_y^0\rceil,0)\\ \quad\vdots\\ C^{110}=(\lceil\boldsymbol{R}_x^1\rceil,\lfloor\boldsymbol{R}_y^1\rfloor,1)\\ C^{111}=(\lceil\boldsymbol{R}_x^1\rceil,\lceil\boldsymbol{R}_y^1\rceil,1)\end{cases} \tag{7}$$

其中:$\lceil\cdot\rceil$、$\lfloor\cdot\rfloor$分别表示对坐标上下取整,$z=0$表示前参考帧时序,$z=1$则表示后参考帧时序。每个顶点对应的三线性重采样权重$\boldsymbol{W}^{ijk}$表示如式(8):

$$\begin{cases}\boldsymbol{W}^{000}=(1-(\boldsymbol{R}_x^0-\lfloor\boldsymbol{R}_x^0\rfloor))(1-(\boldsymbol{R}_y^0-\lfloor\boldsymbol{R}_y^0\rfloor))(1-\Delta z)\\ \boldsymbol{W}^{001}=(1-(\boldsymbol{R}_x^0-\lfloor\boldsymbol{R}_x^0\rfloor))(\boldsymbol{R}_y^0-\lfloor\boldsymbol{R}_y^0\rfloor)(1-\Delta z)\\ \quad\vdots\\ \boldsymbol{W}^{110}=(\boldsymbol{R}_x^1-\lfloor\boldsymbol{R}_x^1\rfloor)(1-(\boldsymbol{R}_y^1-\lfloor\boldsymbol{R}_y^1\rfloor))\Delta z\\ \boldsymbol{W}^{111}=(\boldsymbol{R}_x^1-\lfloor\boldsymbol{R}_x^1\rfloor)(\boldsymbol{R}_y^1-\lfloor\boldsymbol{R}_y^1\rfloor)\Delta z\end{cases} \tag{8}$$

将输入相邻参考帧$(\boldsymbol{I}_{t-1}^*,\boldsymbol{I}_{t+1}^*)$的相关体素记为$\boldsymbol{X}$,最终重采样点$\hat{C}\in\boldsymbol{I}_t^*$的三线性插值方程表示为式(9):

$$\hat{C} = \sum_{i,j,k \in [0,1]} \boldsymbol{W}^{ijk} \boldsymbol{X}(C^{ijk}) \tag{9}$$

重采样点$\hat{C}$的有序集结即为待估计中间帧$\boldsymbol{I}_t^*$。

2 实验设置

2.1 数据准备与实验配置

由于真实的模糊–清晰图像对难以采集，且真实场景的模糊是非线性的[19]。因此在历来的基于深度学习的去模糊方法中，训练数据集常常通过图像退化算法仿真模糊。本文所采用数据集的具体制作方法及实验配置阐述如下。

本文采用了UCF-101公开数据集[20]，该数据集用大量短视频记录了人类各式活动达101类，运动场景及运动种类十分丰富。首先，将数据集中每一段视频抽取为视频帧并将图像分辨率调整为256×256，每连续3帧分为一对，其中，第1帧和第3帧将作为去模糊网络模块训练时的真实值，第2帧将作为插帧网络训练时的真实值。另外，为了保证帧间的运动适宜插帧网络的训练，通过计算图像对帧间的峰值信噪比均值定出上下阈值作为筛选依据。最后，对筛选出来的每一对视频帧卷积以仿真模糊核，模拟模糊程度在一定范围内随机，以增强数据集的随机性并避免模糊过强而不可恢复。最终生成模糊视频数据集含视频帧大约276 860对。

实验的硬件环境为Intel Core i7-7700 CPU 3.60 GHz，显卡NVIDIA GeForce GTX 1070，主机内存16 GB。软件环境为Ubuntu16.04操作系统，CUDA9.0和cuDNN7.1.4。基于深度学习框架TensorFlow，编程语言为Python3.5。网络训练采用Adam优化器[21]并指定参数一阶矩估计的指数衰减率β_1=0.9，二阶矩估计的指数衰减率β_2=0.999。初始化方式为Xavier，批处理个数设置为8。初始学习率设置为1E–4，且每迭代10万次下降至原来的1/10。

2.2 联合训练策略

本文目的是解决模糊视频的插帧问题，集合了去模糊与插帧两种任务，而针对不同任务的两种算法简单组合进行训练不能很好地拟合。因此，为了使整体网络达到更好的收敛，实验一种先分后合，又从粗略到细致(Coarse-to-Fine)的联合训练策略。

首先，分别对两个网络模块进行适当训练。在去模糊模块的网络训练中，通过计算输出值与真实值之间的L2范数损失函数助于网络的反向传播，损失函数设置如式(10)：

$$\mathcal{L}_d = \frac{1}{N}\sum_{i=1}^{N}\left\| \boldsymbol{I}'_{s_i} - \boldsymbol{I}_{s_i} \right\|_2^2 \tag{10}$$

其中：N为批处理个数，$\boldsymbol{I}'_s$为退化前的真实值图像，$\boldsymbol{I}_s$为估计得到的去模糊图像。

在对插帧模块的网络训练中采用L1范数损失函数，设置如式(11)：

$$\mathcal{L}_f = \frac{1}{N}\sum_{i=1}^{N}\left\| \boldsymbol{I}'_{t_i} - \boldsymbol{I}^*_{t_i} \right\|_1 \tag{11}$$

其中：$\boldsymbol{I}'_t$为中间帧的真实值图像，$\boldsymbol{I}_t^*$为插值帧。

图5即展示了两个网络模块分别训练时的收敛情况。随着网络迭代次数增加，模型损失函数值趋于收敛。

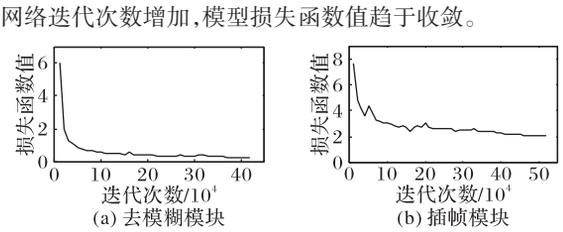

图5 损失函数

Fig. 5 Loss functions

然后，将两个模块的预训练模型导入整体网络进行联合训练。在联合训练阶段，为了增强不同模块参数间的耦合性，将联调网络的损失函数$\mathcal{L}_{\text{total}}$设置如式(12)：

$$\mathcal{L}_{\text{total}} = (1-\lambda)\mathcal{L}_d + \lambda\mathcal{L}_f \tag{12}$$

其中：$0 \leqslant \lambda \leqslant 1$为调整模块间的权重系数。经过反复的实验测试，调节$\lambda$ = 0.1，能够使网络具有更好的图像同步去模糊插帧的能力。

为了使两个模块模型的参数更快速有效地耦合，分为两个阶段进行训练。首先，在联合训练的第一阶段，通过优化器设置去模糊模块中的参数固定不变，单独放开更新插帧模块的网络参数；而后，经过一定次数的迭代进入联合训练的第二阶段，此时再次设置优化器打开去模糊模块的参数更新权限，同时优化网络所有参数，如此细调达到最好效果。

3 实验结果与分析

3.1 客观指标评估

DVF[5]提供了包含379对序列的UCF-101测试集，经过图像退化处理得到最终用于模糊图像插帧的测试集。为了验证联合训练策略的有效性，分别选取了两组前沿的去模糊算法与插帧算法进行简单组合，再与本文经过两个训练阶段的模型结果作对比。分别计算了三种客观评价指标：峰值信噪比(Peak Signal-to-Noise Ratio，PSNR)、结构相似性(Structural SIMilarity，SSIM)以及插值误差(Interpolation Error，IE)[22]。对于参考帧$I(x,y)$和生成帧$\hat{I}(x,y)$，上述指标分别表示为：

$$PSNR(I,\hat{I}) = 10\lg\left(\frac{MAX^2}{MSE(I(x,y),\hat{I}(x,y))}\right) \tag{13}$$

其中，MAX是图像颜色最大值，8位采样点取值255，MSE表示两幅图像间的均方误差。计算所得$PSNR$值越大，代表失真越少。

$$SSIM(I,\hat{I}) = \frac{(2\mu_i\mu_{\hat{i}} + c_1)(2\sigma_{i\hat{i}} + c_2)}{(\mu_i^2 + \mu_{\hat{i}}^2 + c_1)(\sigma_i^2 + \sigma_{\hat{i}}^2 + c_2^2)} \tag{14}$$

其中：μ表示图像的全像素均值，σ表示图像的全像素标准差，$\sigma_{i\hat{i}}$表示图像间的协方差，c为避免被除数为零而设定的常数。SSIM值越接近1，代表图像结构恢复越好。

$$IE(I,\hat{I}) = \left[\frac{1}{N}\sum_{(x,y)}(I(x,y) - \hat{I}(x,y))^2\right]^{\frac{1}{2}} \tag{15}$$

其中N表示图像中所有像素的数量。IE数值越低，代表插值误差越低。

尺度循环网络(Scale-Recurrent Network，SRN)[13]采用了多尺度信息共享的思想取得了不错的去模糊效果。插帧算法方面，CyclicGen[8]运用了循环一致性损失[23]并提取边缘信息以保护图像结构，SuperSlomo[6]则提取融合了帧间的双向光流信息。鉴于上述算法的预训练模型公开且效果可靠，特选取搭配为以下两组：SRN+CyclicGen与SRN+SuperSlomo，以便对比验证本文算法模型的结果。所有测得指标统计如表3。

表3 各模型在UCF-101测试集中的评估结果

Tab. 3 Evaluation of each model on UCF-101 testing set

模型	PSNR/dB	SSIM	IE
SRN+CyclicGen	25.37	0.860	19.93
SRN+SuperSlomo	25.43	0.861	19.81
Ours-stage1	26.31	0.874	18.30
Ours-stage2	**26.84**	**0.881**	**17.82**

由表3数据可以看出,两组简单组合算法的模型测试所得指标较于本文算法的模型的结果显然偏低。对比指标较高的SRN+SuperSlomo,本文方法Ours-stage2所合成中间帧的峰值信噪比提高1.41 dB,结构相似性提升0.020,插值误差降低2.11,说明本文所提出网络结构在去模糊与插帧任务上有着显著作用。而相对的,本文在第一阶段通过单独更新插帧模块参数的联合训练模型(Ours-stage1)在指标上明显有所提升,而联合训练第二阶段最终模型(Ours-stage2)取得最佳指标,其合成帧的峰值信噪比较前者提高0.53 dB,同时结构相似性提升0.007,插值误差降低0.52,说明联调训练策略效果显著。

3.2　主观视觉对比

本节分别对两组前沿算法组合的模型及本文联合训练两个阶段模型输出的中间帧进行视觉对比。

如图6所示,图6(a)是两帧模糊的连续输入帧,内容为少女由前往后转动腰间的呼啦圈,呼啦圈是图像中主要的运动物体,即待估计的关键目标。图6(b)是待估计中间清晰帧的真实值,其整体的图像清晰度高,图像中呼啦圈的位置契合相邻帧的运动趋势。余下四对图像图6(c)~(f)为不同算法模型预测的中间帧图像进行对比分析,它们都一定程度去除了图像模糊并且作出了对输入图像对的中间帧运动估计。

图6　各算法模型的视觉对比

Fig. 6　Visual comparison of each algorithm model

相对的,图6(c)与(d)两组算法的模型所预测图像不能很好地拟合帧间运动变化,导致合成帧中呼啦圈尾部的像素出现大幅错乱。而本文方法的结果则接近于图6(b),对呼啦圈形状的预测更加完整且符合真实世界的运动趋势。相较于联合训练第一阶段模型的结果图6(e),本文第二阶段模型所得结果图6(f)对呼啦圈尾部的预测更加平滑,像素混乱现象微乎其微,整体画面观感和谐。

总的来说,本文所用联合训练策略提升了对模糊视频帧序列中间帧的预测质量,取得更为清晰且运动估计更加合理的重制结果。然而图像模糊经过修复依然不可避免地会丢失一些纹理细节,因此基于重建图像估计所得中间帧在评估指标和视觉结果没有达到特别高的精度,这也是不得不面对的挑战。

3.3　重制序列展示

本节进一步展示了通过本文方法取得的三组重制序列样例。如图7所示,对于每一组样例,前两列为输入的两帧连续模糊帧,后三列为三帧清晰且视觉连贯的重制帧序列。

由图7可见,三组不同模糊程度的两帧输入帧经过本文方法进行重制之后,输出的视频帧序列画面中的人与物都转变清晰,且中间合成帧在内容上符合其前后帧之间的运动趋势,使得源序列中目标动作更加清晰连贯。方框区域标记了组图帧间目标的主要运动变化,样例一(a)中理发师手中的剪刀逐渐抬升;样例二(b)中化妆女士的眼线笔从眼睛中部缓慢移动到眼角;样例三(c)中吹头发女士的眼睛慢慢闭合。上述结果验证了本文去模糊高帧率重制方法的可靠性与适用性。

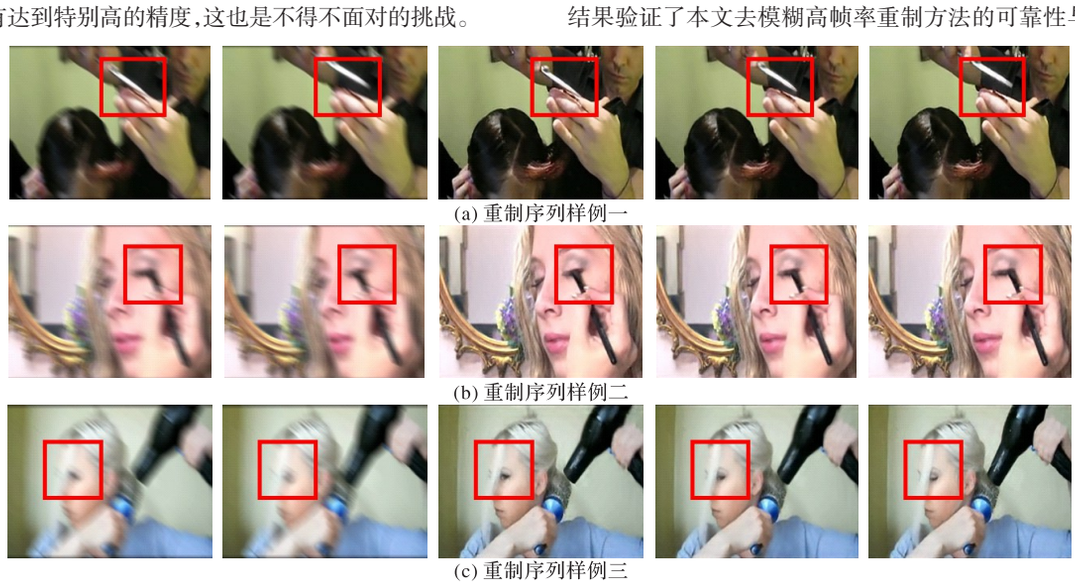

图7　模糊视频帧序列重制结果

Fig. 7　Reconstruction results of blurred video frame sequences

4 结语

提出了一种针对模糊视频的插帧方法。设计了多任务融合的网络模型,制作了大型的模糊视频数据集,利用联合训练策略使网络有效收敛。对比了多种模型,测试指标显著提升,展示了模糊视频帧序列经过重制取得的良好视效,证明本文提出的同步去模糊高帧率方法的可行性,对未来端对端解决多种问题混合的计算机视觉任务具有一定的参考意义。

下一步研究方向大致如下,将去模糊与插帧两个过程中的图像特征提取与信息处理方法进行合理融合以简化整体网络、设计更加合理高效的损失函数、加快深度网络的速度等。

参考文献(References)

[1] DOSOVITSKIY A, FISCHER P, ILG E, et al. FlowNet: learning optical flow with convolutional networks[C]// Proceedings of the 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2015: 2758-2766.

[2] ILG E, MAYER N, SAIKIA T, et al. FlowNet 2.0: evolution of optical flow estimation with deep networks[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 1647-1655.

[3] NIKLAUS S, MAI L, LIU F. Video frame interpolation via adaptive convolution [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 2270-2279.

[4] NIKLAUS S, MAI L, LIU F. Video frame interpolation via adaptive separable convolution[C]// Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 261-270.

[5] LIU Z, YEH R A, TANG X, et al. Video frame synthesis using deep voxel flow[C]// Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 4473-4481.

[6] JIANG H, SUN D, JAMPANI V, et al. Super SloMo: high quality estimation of multiple intermediate frames for video interpolation [C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 9000-9008.

[7] NIKLAUS S, LIU F. Context-aware synthesis for video frame interpolation[C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 1701-1710.

[8] LIU Y L, LIAO Y T, LIN Y Y, et al. Deep video frame interpolation using cyclic frame generation[C]// Proceedings of the 2019 AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2019: 8794-8802.

[9] BAO W, LAI W-S, MA C, et al. Depth-aware video frame interpolation [EB/OL]. [2019-05-23]. https://arxiv.org/pdf/1904.00830.pdf.

[10] WANG R, TAO D. Recent progress in image deblurring[EB/OL]. [2019-05-23]. https://arxiv.org/pdf/1409.6838.pdf.

[11] KUPYN O, BUDZAN V, MYKHAILYCH M, et al. DeblurGAN: blind motion deblurring using conditional adversarial networks [C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 8183-8192.

[12] NAH S, KIM T H, MU LEE K. Deep multi-scale convolutional neural network for dynamic scene deblurring[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 3883-3891.

[13] TAO X, GAO H, SHEN X, et al. Scale-recurrent network for deep image deblurring[C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 8174-8182.

[14] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]// Proceedings of the 2015 International Conference on Medical Image Computing and Computer-Assisted Intervention, LNCS 9351. Cham: Springer, 2015: 234-241.

[15] TAO X, GAO H, LIAO R, et al. Detail-revealing deep video super-resolution[C]// Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 4482-4490.

[16] PATHAK D, KRÄHENBÜHL P, DONAHUE J, et al. Context encoders: feature learning by inpainting[C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 2536-2544.

[17] SIMO-SERRA E, IIZUKA S, SASAKI K, et al. Learning to simplify: fully convolutional networks for rough sketch cleanup[J]. ACM Transactions on Graphics, 2016, 35(4): No. 121.

[18] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778.

[19] SU S, DELBRACIO M, WANG J, et al. Deep video deblurring for hand-held cameras[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 237-246.

[20] SOOMRO K, ZAMIR A R, SHAH M. UCF101: a dataset of 101 human actions classes from videos in the wild[EB/OL]. [2019-05-23]. https://arxiv.org/pdf/1212.0402.pdf.

[21] KINGMA D P, BA J L. Adam: a method for stochastic optimization[EB/OL]. [2019-05-23]. https://arxiv.org/pdf/1412.6980.pdf.

[22] BAKER S, SCHARSTEIN D, LEWIS J P, et al. A database and evaluation methodology for optical flow[J]. International Journal of Computer Vision, 2011, 92(1): 1-31.

[23] KIM T, CHA M, KIM H, et al. Learning to discover cross-domain relations with generative adversarial networks[C]// Proceedings of the 34th Internat*ional* Conference on Machine Learning. New York: JMLR.org, 2017: 1857-1865.

This work is partially supported by the National Natural Science Foundation of China (61802065).

LIN Chuanjian, born in 1994, M. S. candidate. His research interests include computer vision, image processing.

DENG Wei, born in 1992, M. S. His research interests include computer vision, image processing.

TONG Tong, born in 1986, Ph. D., research fellow. His research interests include artificial intelligence and computer vision, medical image processing and analysis.

GAO Qinquan, born in 1986, Ph. D., associate research fellow. His research interests include artificial intelligence and computer vision, medical image processing and analysis, computer-aided surgical guidance.