文章编号：2095-2163(2019)04-0252-07 中图分类号：TP391.4 文献标志码：A

基于深度学习的视频插帧算法

张 倩，姜 峰
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：视频帧率转换技术是利用视频中相邻两帧之间的相关信息并应用插值的方法将中间帧重建出来的一种技术。由于该技术能在编码中去除冗余信息并降低视频传输过程中的帧率，减少视频网络传输的数据量，因此可应用于视频压缩或增强视频连续性。本文将传统方法中的光流估计与深度学习相结合，提出了一种将运动估计和遮挡处理联合建模的视频帧插值的端到端卷积神经网络模型。首先使用改进的 GridNet 网络模型计算输入图像之间的双向光流，根据估计到的双向光流信息与输入图像进行 warp 操作得到 2 个翘曲图像，为解决遮挡问题，使用另一个 GridNet 网络模型重新估计图像的双向光流信息并预测插值帧的像素的可见性，最后将估计到的信息与原图像通过线性融合以形成中间帧。本文还尝试了多种损失函数，最终确定了将 L1 损失、感知损失、warp 损失、平滑度损失等多种损失函数加权而成的损失函数。实验结果证明，本文提出的视频插帧网络结构可以有效提高光流估计的质量并改善遮挡问题，可以生成逼真、自然、质量更好的中间帧。

关键词：视频插帧；深度学习；光流估计

Video interpolation based on deep learing

ZHANG Qian, JIANG Feng
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 The video frame rate conversion technique is a technique for reconstructing an intermediate frame by using correlation information between adjacent two frames in the video and applying interpolation. Since the technology can remove redundant information in encoding and reduce the frame rate during video transmission, and reduce the amount of data transmitted by the video network, it can be applied to video compression or enhance video continuity.In this paper, the optical flow estimation and the deep learning in the traditional method are combined, and an end-to-end convolutional neural network model is proposed, which combines motion estimation and occlusion processing. First, the improved GridNet network model is used to calculate the bidirectional optical flow between the input images, and two warped images are obtained based on warp operation of the estimated bidirectional optical flow information and the input image. After that to solve the occlusion problem,the paper uses another GridNet network. The model re-estimates the bidirectional optical flow information of the image and predicts the visibility of the pixels of the interpolated frame, and linearly fuses the estimated information with the original image to form an intermediate frame.Finally, a variety of loss functions are also tried, and the loss function is determined that weighted various loss functions such as L1 loss, perceptual loss, wrap loss, and smoothness loss. The experimental results show that the video frame interpolation network structure proposed in this paper can effectively improve the quality of optical flow estimation and improve the occlusion problem, and can generate intermediate frames with vividness and naturality and better quality.

【Key words】 frame interpolation; deep learning; optical flow estimation

0 引 言

伴随智能终端及多媒体技术的迅猛发展，视频应用更加多样化，其中涉及到的视频内容和种类正陆续增多。与此同时，高清晰度的显示设备也呈现出大规模增长态势，高刷新频率的显示器不断普及，人们对视频分辨率的要求也越来越高。目前情况下，很多视频的帧率通常只用 30 帧/秒，人们在观看这种视频时视觉感知上会出现卡顿等问题，也无法发挥高刷新频率显示器的优势。因此可以将低帧率的视频通过视频帧率转换技术插值为高帧率视频，例如可将帧率为 30 帧/秒的视频提升至帧率为 60 帧/秒，使视频更加地平滑和连续，从而提升人们在观看视频时的逼真度和交互感。

研究可知，作为数据量非常巨大的信息载体，视频在网络传输过程中对带宽的要求非常高，而且存储视频的成本也变得巨大，因此就必须采取高效的视频压缩方法，尽可能去除视频中的冗余成分。主流的视频编码标准 H.265/HEVC[1] 已经在很大程度上降低了视频的冗余信息。视频冗余信息整体上可

作者简介：张 倩(1995-)，女，硕士研究生，主要研究方向：计算机视觉、图像处理、机器学习等；姜 峰(1978-)，男，博士，教授，博士生导师，主要研究方向：图像处理、视频编解码、计算机视觉等。

收稿日期：2019-05-30

分为时间冗余和空间冗余等。其中，时间冗余信息主要指视频相邻帧之间具有相似性，H.265/HEVC 采用帧间预测的方法来去除时间冗余信息，帧间预测则通过将已编码的视频帧作为当前帧的参考，进行运动估计来获取运动信息，从而去除时间冗余。空间冗余信息主要指视频单帧图像在空间上的局部相似性，H.265/HEVC 通常采用帧内预测和变换编码的技术去除空间冗余信息，帧内预测通过已编码的像素预测当前像素去除空间冗余信息。变换编码将图像能量在空间域的分散分布转换至变换域的集中分布，从而去除空间冗余。

尽管最新的国际视频压缩标准 H.265/HEVC 较 H.264/AVC 相比性能上有了显著的提高，可以大幅度地去除视频中的冗余信息。但在网络带宽有限等情况下，这些压缩标准仍然不能满足人们的需求，所以一些研究人员开始尝试用其他的方法手段继续对视频进行压缩处理，其中效果较好的方法为帧率转换技术。视频帧率转换技术是指利用视频中相邻两帧之间的相关信息并应用插值的方法将中间帧重建出来的一种技术。该技术在视频编码中去除冗余信息并降低视频传输过程中的帧率，有效减少视频网络传输的数据量。视频插帧技术的效果决定了重建帧的质量，因此视频插帧技术在视频压缩领域有着重要影响。

视频插帧技术一直是计算机视觉领域的热点研究内容之一，视频插帧技术的提升对帧率转换技术及视频压缩技术的研发起着举足轻重的作用，同时视频插帧技术也广泛应用于慢镜头回放等场景中，这也推动相关研究人员不断地改进这一技术，并积极展开更深层次的探索。

1　相关工作

随着深度学习在计算机领域不断取得成功，研究者们即尝试将深度学习与视频插帧技术相结合来满足插帧需求。视频插帧技术是指利用视频中相邻前后帧之间的相关信息，应用插值的方法获得中间帧。根据新的插值帧的数量与输入视频帧的数量关系，视频插帧可分为均匀插帧与非均匀插帧，如图 1 所示。相应地，均匀插帧是指新的插值帧与输入的视频帧序列按照 1∶1 的比例合成新的视频序列，非均匀插帧一般是指新的插值帧与输入的视频序列按照如图 1 所示 2∶3 的比例合成新的视频序列，本文着重研究的是均匀的视频插帧技术。

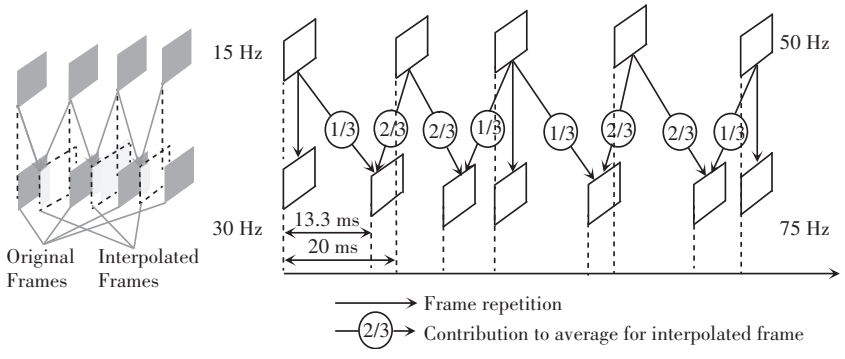


图 1　均匀插帧与非均匀插帧示意图

Fig. 1　Uniform frame insertion and non-uniform frame insertion

本文拟要研究的是基于深度学习的视频插帧技术。研究中，是将传统方法中的光流估计与深度学习相结合，提出了一种将运动估计和遮挡处理联合建模的视频帧插值的端到端卷积神经网络模型。文中首先使用 GridNet 卷积神经网络模型估计输入的相邻帧之间的正反双向光流。受增强学习的启发，为了将光流估计的差异向下传递，研究将估计到的正反向光流与输入的相邻帧图像进行 warp 操作获得 2 个图像，将 warp 得到的图像与原始输入图像及预测到的双向光流共同作为下一个卷积神经网络的输入。为解决光流估计中的遮挡问题，研究中同时使用另一个 GridNet 卷积神经网络模型预测像素的可见性，并重新预测光流信息。最后则将预测到的光流信息及可见性因子等通过线性融合来形成视频插帧过程中的中间帧。

2　算法详述

2.1　算法原理

给定 2 个输入图像 I_0 和 I_1 以及时间 $t \in (0,1)$，本文目标是在 $T = t$ 时预测中间图像 $\hat{I}_t$。一种较为直接的方法是，训练神经网络[2]直接输出 $\hat{I}_t$ 的 RGB 像素。然而，为了做到这一点，网络不仅必须学习运动补偿，而且还要学习 2 个输入图像的外观。由于

RGB 色彩空间丰富,以这种方式很难生成高质量的中间图像。受文献[3-6]最新进展的启发,本文尝试在 $T = t$ 时刻融合扭曲的输入图像。让 $F_{t\to 0}$ 和 $F_{t\to 1}$ 分别表示从 I_t 到 I_0 和 I_t 到 I_1 的光流。如果已知这 2 个流,就可以合成中间图像 $\hat{I}_t$,其数学计算公式为:

$$\hat{I}_t = \alpha \otimes f_{bw}(I_0, F_{t\to 0}) + (1-\alpha) \otimes f_{bw}(I_1, F_{t\to 1}), \tag{1}$$

其中,$f_{bw}(,)$ 是一个后向扭曲函数,可以使用双线性插值[7]实现,并且是可微分的;参数 α 控制 2 个输入图像的权值,取决于 2 个因素:时间一致性和遮挡预测;$\otimes$ 表示元素乘法,意味着输入图像的内容感知加权。对于时间一致性,$T = t$ 的时间步长越接近 $T = 0$,I_0 对 $\hat{I}_t$ 的权值就越大,反之 I_1 也有类似的情况。

另据研究可知,视频帧插值问题的一个重要特性是,如果像素 p 在 $T = t$ 处可见,则很可能在输入图像之一中至少可见,这意味着可以解决遮挡问题。因此,研究引入了可见性图 $V_{t\leftarrow 0}$ 和 $V_{t\leftarrow 1}$。$V_{t\leftarrow 0}(p) \in [0,1]$ 表示当从 $T=0$ 移动到 $T=t$ 时,像素 p 是否保持可见(0 完全被遮挡)。结合时间一致性和遮挡预测,研究推得:

$$\hat{I}_t = \frac{1}{Z} \otimes ((1-t) V_{t\leftarrow 0} \otimes f_{bw}(I_0, F_{t\to 0}) + t V_{t\leftarrow 1} \otimes f_{bw}(I_1, F_{t\to 1})), \tag{2}$$

其中,$Z = (1-t)V_{t\to 0} + tV_{t\to 1}$ 是一个归一化因子。

考虑到研究无法访问目标中间图像 I_t,因此很难计算光流 $F_{t\to 0}$ 和 $F_{t\to 1}$。针对这一问题,若要预测光流 $F_{t\to 0}$ 和 $F_{t\to 1}$,可以使用 2 个输入图像 $F_{0\to 1}$ 和 $F_{1\to 0}$ 之间的光流大约合成中间光流。

假设光流场是局部平滑的。具体来说,$F_{t\to 1}(p)$ 可以近似为:

$$\hat{F}_{t\to 1}(p) = -(1-t) F_{1\to 0}(p), \tag{3}$$

研究在相同或相反的方向上取 2 个输入图像之间的光流方向,并相应地调整公式(3)的幅度 $(1-t)$。类似于 RGB 图像合成的时间一致性,可以通过如下组合双向输入光流来近似中间光流(以矢量形式表示)。可参考写作如下数学形式:

$$\hat{F}_{t\to 1} = (1-t)^2 F_{0\to 1} - t(1-t) F_{1\to 0}, \tag{4}$$

这种近似在平滑区域中表现良好,但在运动边界周围很差,因为运动边界附近的运动不是局部平滑的。为了减少可能导致图像合成不良的运动边界周围的伪影,实验再次学习近似光流。过程中借鉴光流估计的级联架构的思想,研究训练了一个光流插值子网络。该子网获取输入图像 I_0 和 I_1,两者之间的光流 $F_{0\to 1}$ 和 $F_{1\to 0}$,近似光流 $\hat{F}_{t\to 0}$ 和 $\hat{F}_{t\to 1}$,以及使用近似光流 warp 操作生成的 $f_{bw}(I_0, \hat{F}_{t\to 0})$ 和 $f_{bw}(I_1, \hat{F}_{t\to 1})$ 两个变形图像作为输入,并输出重新学习预测的中间光流场 $F_{t\to 0}$ 和 $F_{t\to 1}$。

如前文所述,可见性图对于处理遮挡是必不可少的。因此,研究还使用光流插值 CNN 网络预测 2 个可见性图 $V_{t\leftarrow 0}$ 和 $V_{t\leftarrow 1}$,并强制其满足以下约束:

$$V_{t\leftarrow 0} = 1 - V_{t\leftarrow 1}. \tag{5}$$

没有这样的约束,网络培训就会发生分歧。遮挡问题的直观表示见图 2。由图 2 分析可知,$V_{t\leftarrow 0}(p) = 1$ 意味着 $V_{t\leftarrow 1}(p) = 0$,意味着 I_0 中的像素 p 在 $T = t$ 被遮挡,因此应该完全信任 I_1,反之亦然。还需注意,很少发生时间 t 处的像素在时间 0 和 1 处都被遮挡。预测可见性图的样本最好以彩色显示,其中 $t = 0.5$。手臂从 $T = 0$ 向下移动到 $T = 1$。因此 $T = 0$ 手臂右上方的区域在 t 处可见,但 $T = 1$ 右侧上方的区域在 t 处被遮挡、即不可见。下方的可见性图清楚地显示了这种现象。$V_{t\leftarrow 0}$ 中臂周围的区域表示 I_0 中的这些像素对合成的 $\hat{I}_t$ 贡献最大,而 I_1 中的遮挡像素贡献很少。类似的现象也发生在运动边界周围(例如,运动员的身体周围)。

(a) Frame 0 (b) Frame 1

(c) Framet_groundtruth (d) Framet_ours

图 2 直观表示遮挡问题

Fig. 2 Visual representation of occlusion problems

2.2 网络结构

研究中用到的网络结构整体如图 3 所示。对于

光流计算和光流插值 CNN，研究采用 GridNet 架构[8]。GridNet 是一个完全卷积的神经网络，由编码器和解码器组成，对于 2 个网络，在相同的空间分辨率下将编码器和解码器进行连接。研究中的编码器共有 6 个层次结构，包括 2 个卷积和 1 个 Leaky ReLU（$\alpha = 0.1$）图层。设计时，除去最底层之外的每个层次结构的末尾，使用步幅为 2 的平均池化层来减小空间维度。解码器部分配置了 5 个层次结构。在每个层次结构的起始处，使用双线性上采样层将空间维度增加 2 倍，紧接着就是 2 个卷积和 Leaky ReLU 层。

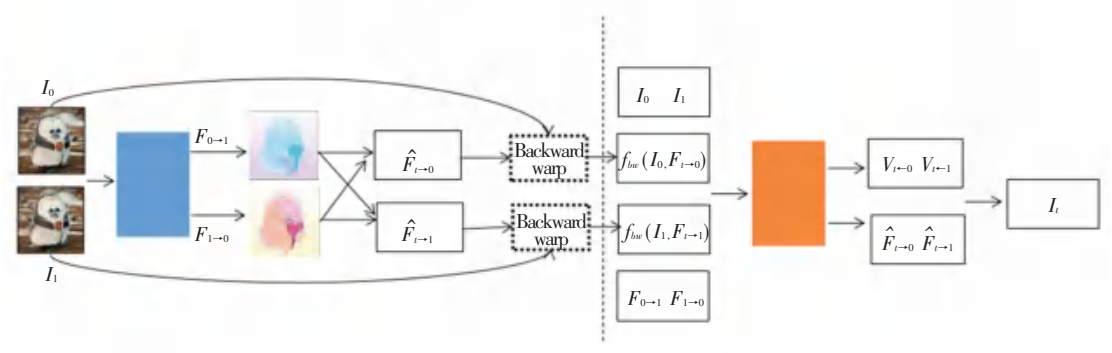


图 3　网络结构图

Fig. 3　Network structure diagram

对于光流计算 CNN，在编码器的前面数层中使用大型滤波器以捕获大幅运动是至关重要的。因此，研究在前两个卷积层中使用 7 * 7 内核，在第二层次中使用 5 * 5。对于整个网络其余部分的层，研究使用了 3 * 3 卷积内核。

2.3　学习算法

本文中的网络使用 Adam 优化器[9]训练 500 次迭代。学习率初始化为 0.000 1，每 200 个迭代减少 10 倍。在训练期间，所有视频剪辑先被分成较短的视频剪辑，每个视频剪辑中有 12 帧，并且 2 个剪辑中的任何一个之间没有重叠。对于数据增强，研究将随机反转整个序列的方向，并选择 9 个连续帧进行训练。

2.4　损失函数

对于视频插帧合成的插值帧的质量良好与否，损失函数起着不可低估的作用。研究时最直接的衡量模型效果的损失函数就是计算合成帧与真实帧之间的像素误差，这种方法虽然可以得到高优的量化指标，但是人眼往往对这种像素级别的微小误差并不敏感，而是更加关注图像的边缘及纹理信息。广泛调研后可知，并未发现哪种损失函数的性能堪称完美，而是各占胜场、也各有不足，因此可将多种损失函数进行综合加权，这样一来也许会取得较好效果。此处，给定输入图像 I_0 和 I_1，两者之间有一组中间帧 $I_t(t \in (0,1))$。文中针对实验时采用的 4 种损失函数，可做研究阐析论述如下。

2.4.1　L1 范数与 L2 范数

L1 范数与 L2 范数直接体现了像素之间的误差，是计算像素级别误差最直接的损失函数。在实验中，L1 范数与 L2 范数展现出了类似的特性，但有相关实验表明，在处理图像问题时，使用 L1 范数会比使用 L2 范数得到更加清晰的图像结果，因此本文拟使用 L1 范数作为损失函数之一。假设真实的中间帧为 I_t，生成的中间帧为 $\hat{I}_t$，则 L1 损失为：

$$l_r = \| \hat{I}_t - I_t \|_1. \tag{6}$$

像素之间的误差小并不代表着肉眼对 2 张图片感受相同，由于自然图像遵循多模态分布，因此 L1 范数与 L2 范数收敛的结果非常模糊。

2.4.2　感知损失

感知损失在文献[10]中首获提出，用于图像风格转换。感知损失如图 4 所示。

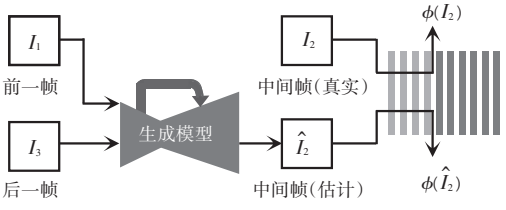


图 4　感知损失

Fig. 4　Perceived loss

感知损失将已经训练好的 VGG-16[11]网络作为损失函数的一部分。研究可知，感知损失是从预先训练的 VGG 网络中提取特征误差、而非像素误

差,因此就包含了对高频细节的感受能力,这也是 L1 范数与 L2 范数所不具备的。在实际训练中,加入了感知损失后,纹理和细节得到了明显的增强。

由于本课题的任务与风格转换不同,更加需要逼近真实的图像,因而即需使用更加浅层的网络输出。经过调试,使用 VGG-16 的第二层池化层作为输出获得了良好效果。感知损失的计算公式如下:

$$l_p = \| \phi(\hat{I}_t) - \phi(I_t) \|. \quad (7)$$

其中,ϕ 函数为 ImageNet 预训练 VGG16 模型的第二层池化层之前的网络结构,包括 4 个卷积层和 2 个池化层。

2.4.3 warp 损失

warp 损失主要用来计算光流估计的质量。光流信息则是用于表示相邻帧对应像素位置的运动矢量信息,而 warp 操作可依次分为 2 个步骤:像素映射以及二维线性插值,因此当已知相邻帧之间的光流场信息时,就可以通过变形操作将一帧图像映射为另一帧图像。再通过计算目标图像与变形操作合成的新图像间的差异,就可以用来检测光流估计的质量。warp 损失的计算公式如下:

$$l_w = \| I_0 - f_{bw}(I_1, F_{0\to1}) \|_1 + \| I_1 - f_{bw}(I_0, F_{1\to0}) \|_1 + \| I_t - f_{bw}(I_0, \hat{F}_{t\to0}) \|_1 + \| I_t - f_{bw}(I_1, \hat{F}_{t\to1}) \|_1. \quad (8)$$

2.4.4 平滑度损失

平滑度损失的计算公式如下:

$$l_s = \| \tilde{\nabla} F_{0\to1} \|_1 + \| \tilde{\nabla} F_{1\to0} \|_1, \quad (9)$$

平滑度损失与 warp 损失相结合都是用来检测光流估计的质量。

由于本文所讨论的 L1 范数与 L2 范数、感知损失、warp 损失、平滑度损失单独应用到视频插帧技术中都存在一定的缺陷与不足。其中,L1 范数和 L2 范数直接体现了像素级别的误差,训练速度快,但是却只是给出了像素的差值而忽略了对插值帧图像结构等差异的计算。感知损失虽然有效解决了使用像素误差产生的模糊问题,更加关注了图像的细节纹理信息,但是感知损失对图像的低频信息却并不敏感。单独使用 warp 损失和平滑度损失仅能衡量光流估计的质量而无法准确衡量最终插值结果的质量。因此经过仿真验证,将多种损失函数引入加权处理会弥补单独使用损失函数的缺点,将会取得比较良好的结果。给定输入图像 I_0 和 I_1,两者之间有一组中间帧 $I_t(t \in (0,1))$,本次研究的最终损失函数是 4 种损失函数的线性组合,其数学公式可表示为:

$$L = \lambda_r l_r + \lambda_p l_p + \lambda_w l_w + \lambda_s l_s. \quad (10)$$

其中,权重 $\lambda_r = 0.8$,$\lambda_p = 0.005$,$\lambda_w = 0.4$ 和 $\lambda_s = 1$ 凭经验设置。

同时,文中网络的每个组成部分都是可区分的,包括 warp 和光流计算。因此,实验的模型可以进行端到端的训练。

3 实验

3.1 实验参数及数据集

本文使用 Adam 学习算法优化训练模型,进行参数更新,设置 Adam 学习算法中相关参数为:$\alpha = 0.000\,1$,$\beta_1 = 0.9$,$\beta_2 = 0.999$,$\varepsilon = 10^{-8}$。

研究中训练使用的数据集是从 YouTube 收集的 240-fps 视频,数据集中都有各种各样的场景,从室内到室外、从静态到移动的摄像机、从日常活动到专业运动等。在训练期间,所有视频剪辑先被分成较短的视频剪辑,每个视频剪辑中有 12 帧,并且 2 个剪辑中的任何一个之间没有重叠。对于数据增强,研究中随机反转整个序列的方向,并选择 9 个连续帧进行训练。在图像级别上,每个视频帧的大小被调整为较短的空间维度,再结合水平移动随机裁剪为 352 * 352 大小。

3.2 实验结果

研究使用所有数据训练本文组建的网络,并在多个独立的数据集上测试本文的模型,包括 Middlebury 基准、UCF101、慢流数据集和高帧率 Sintel 序列。

总地来说,对于 Middlebury,研究中将 8 个序列的单帧视频插值结果提交给其评估服务器。对于 UCF101,在每三帧中,第一和第三帧用作输入,预测第二帧。慢速流动数据集包含 46 个使用专业高速摄像机拍摄的视频。研究使用第一和第三视频帧作为输入,并插入中间帧,相当于将 30-fps 视频转换为 60-fps 视频。

最初的 Sintel 序列以 24-fps 渲染。其中 13 个以1 008-fps重新渲染。要使用视频帧插值方法从 24-fps 转换为 1008-fps,需要插入 41 个中间帧。然而,正如在前文中所分析的那样,使用递归单帧插值方法不能直接实现这一点。因此,本次研究预测 31 个中间帧,以便与先前的方法做出公平比较。

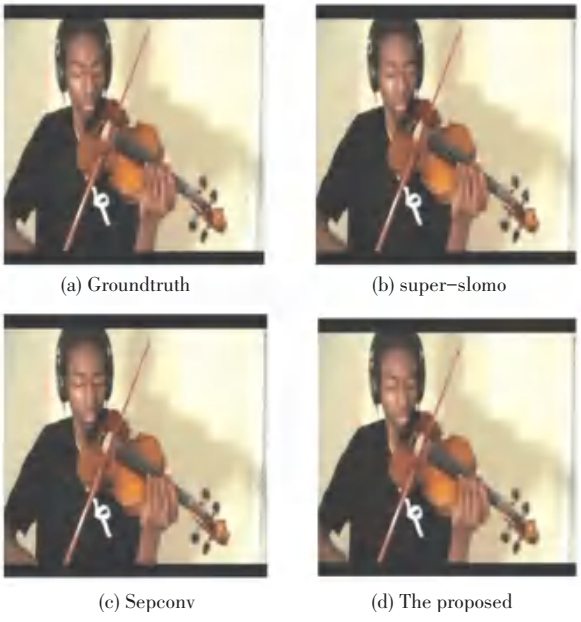
实验中,研究对比了训练样本数量的影响,这里比较 2 个模型。一个仅在 Adobe240-fps 上训练,另一个在完整的数据集上进行训练,2 个模型在 UCF101 数据集上的性能见表 1。

表 1　UCF101 数据集结果

Tab. 1　The results of the UCF101 dataset

方法	UCF-101	
	PSNR	*SSIM*
DVF	32.46	0.930
FlowNet2	32.30	0.930
SepConv	33.02	0.935
Ours-240fps	32.84	0.935
The proposed	**32.91**	**0.938**

从表 1 的实验结果可以看出，训练数据越多，本文模型的效果越好。以图片的形式给出对比效果如图 5 所示。网络模型在慢数据集下的实验结果见表 2。网络模型在高帧率数据集下的实验结果见表 3。



(a) Groundtruth　(b) super-slomo

(c) Sepconv　(d) The proposed

图 5　UCF101 实验结果

Fig. 5　The result of the UCF101 dataset

表 2　慢数据集结果

Tab. 2　The results of slowflow dataset

方法	*PSNR*	*SSIM*
Phase-Based	31.05	0.858
FlowNet2	33.30	0.930
SepConv	31.79	0.895
The proposed	33.91	0.928

表 3　高帧率 Sintel 数据集结果

Tab. 3　The results of the high-frame-rate Sintel dataset

方法	*PSNR*	*SSIM*
Phase-Based	28.75	0.840
FlowNet2	30.30	0.923
SepConv	31.79	0.907
The proposed	32.27	0.927

在本节中，研究将本文的方法与最先进的方法进行比较，包括基于相位的插值、可分离的自适应卷积(SepConv)和深度体素流(DVF)。实验表明，本文提出的网络模型可以获得更加精确的视频插帧效果。

4　结束语

本文研发提出了一种端到端可训练的 CNN，可以在 2 个输入图像之间根据需要产生尽可能多的中间视频帧。首先使用流量计算 CNN 来估计 2 个输入帧之间的双向光流，并且 2 个流场线性融合以接近中间光流场。然后，使用流动插值 CNN 来重新定义近似流场并预测用于插值的软可见性图。接下来，又使用超过 1.1K 240-fps 的视频剪辑来训练本文的网络预测 7 个中间帧。对单独验证集的消融研究证明了流动插值和可见性图的优势。仿真实验证明，本文的多帧方法在 Middlebury、UCF101、慢速流和高帧率 Sintel 数据集上始终优于最先进的单帧方法。对于光学流的无监督学习，本文研发的网络也要优于 KITTI 2012 基准测试中最近的 DVF 方法。

参考文献

[1] SULLIVAN G J, OHM J R, HAN W J, et al. Overview of the high efficiency video coding (HEVC) standard [J]. IEEE Transactions on circuits and systems for video technology, 2012, 22(12):1649-1668.

[2] LONG G, KNEIP L, ALVAREZ J M, et al. Learning image matching by simply watching video[M]//LEIBE B, MATAS J, SEBE N, et al. Computer Vision - ECCV 2016. ECCV 2016. Lecture Notes in Computer Science. Cham: Springer, 2016, 9910: 434-450.

[3] BAKER S, SCHARSTEIN D, LEWIS J P, et al. A database and evaluation methodology for optical flow [C]//2007 IEEE 11th International Conference on Computer Vision. Rio de Janeiro, Brazil: IEEE, 2007: 1-8.

[4] NIKLAUS S, MAI Long, LIU Feng. Video frame interpolation via adaptive convolution [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 2270-2279.

[5] NIKLAUS S, MAI Long , LIU Feng. Video frame interpolation via adaptive separable convolution [C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 261-270.

[6] LIU Ziwei, YEH R, TANG Xiaoou, et al. Video frame synthesis using deep voxel flow [C]//Proceedings of International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 1-10.

[7] ZHOU Tinghui, TULSIANI S, SUN Weilun, et al. View synthesis by appearance flow[C]//LEIBE B, MATAS J, SEBE N, et al. Computer Vision - ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, Cham: Springer, 2016, 9908: 286-301.

(下转第 262 页)