

用于视频修复的连贯语义时空注意力网络

刘浪 李梁 但远宏

重庆理工大学计算机科学与工程学院 重庆 400054

(lang224017@gmail.com)



摘要 现有的视频修复方法通常会产生纹理模糊、结构扭曲的内容以及伪影,而将基于图像的修复模型直接应用于视频修复会导致时间上的不一致。从时间角度出发,提出了一种新的用于视频修复的连贯语义时空注意力(Coherent Semantic Spatial-Temporal Attention,CSSTA)网络,通过注意力层,使得模型关注于目标帧被遮挡而相邻帧可见的信息,以获取可见的内容来填充目标帧的孔区域(hole region)。CSSTA层不仅可以对孔特征之间的语义相关性进行建模,还能对远距离信息和孔区域之间的远程关联进行建模。为合成语义连贯的孔区域,提出了一种新的损失函数特征损失(Feature Loss)以取代 VGG Loss。模型建立在一个双阶段粗到精的编码器-解码器结构上,用于从相邻帧中收集和提炼信息。在 YouTube-VOS 和 DAVIS 数据集上的实验结果表明,所提方法几乎实时运行,并且在修复结果、峰值信噪比(PSNR)和结构相似度(SSIM)3个方面均优于3种代表性视频修复方法。

关键词: 视频修复;图像修复;时空注意力;特征损失;VGG Loss

中图法分类号 TP391.4

Coherent Semantic Spatial-Temporal Attention Network for Video Inpainting

LIU Lang, LI Liang and DAN Yuan-hong

College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China

Abstract Existing video inpainting methods usually produce blurred texture, distorted structure and artifacts, while applying image-based inpainting model directly to the video inpainting will lead to inconsistent time. From the perspective of time, a novel coherent semantic spatial-temporal attention(CSSTA) for video inpainting is proposed, through the attention layer, the model focuses on the information that the target frame is partially blocked and the adjacent frames are visible, so as to obtain the visible content to fill the hole region of the target frame. The CSSTA layer can not only model the semantic correlation between hole features but also remotely correlate the long-range information with the hole regions. In order to complete semantically coherent hole regions, a novel loss function Feature Loss is proposed to replace VGG Loss. The model is built on a two-stage coarse-to-fine encoder-decoder model for collecting and refining information from adjacent frames. Experimental results on the YouTube-VOS and DAVIS datasets show that the method in this paper runs almost in real-time and outperforms the three typical video inpainting methods in terms of inpainting results, peak signal-to-noise ratio (PSNR) and structural similarity (SSIM).

Keywords Video inpainting, Image inpainting, Spatial-Temporal attention, Feature Loss, VGG Loss

1 引言

视频修复的目标是使用语义连贯和时间一致的内容来填充给定视频序列的缺失区域(孔区域)。视频修复(也称为视频合成)有许多现实世界的应用,如电影后期处理、不需要的对象移除和视频恢复。同时,视频修复可以与增强现实(AR)结合使用,以提供更好的视觉体验。目前,视频修复方法主要分为两大类,一类是基于图像修复逐帧进行修复的方法,另一类是直接对整个视频进行修复的方法。

图像修复方法可分为非学习的方法^[1-8]和学习的方法^[9-16]。非学习的方法把图像分成小块,并通过将最相似的块粘贴到图像中的某个位置来恢复掩码区域;基于学习的方法主要利用深度生成神经网络直接进行图像修复,上下文编码器^[11]用于负责图像修补任务的深度神经网络。Yang等^[12]将对抗性训练引入新颖的编码器-解码器管道,并输出缺失区域的预测。此后不久,Iizuka等^[9]扩展了这项工作,并提出了局部和全局的判别器,以提高修复质量,该方法需要进行后期处理,以增强孔边界附近颜色的一致性。Yu等^[14]进一步提出使用门控卷积自动学习掩码,并与 SN-PatchGAN 判别器结合以实现更好的预测。将视频拆分为逐帧图像进行修复,

到稿日期:2020-06-20 返修日期:2021-02-04 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。

基金项目:国防科技创新特区项目

This work was supported by the National Defense Technology Innovation Zone Project.

通信作者:但远宏(dyh@cqu.edu.cn)

忽略了视频动态的运动规律,难以预测图像空间随时间的位置变化,由此导致时间上的不一致并产生了明显的闪烁伪影,如图 1 所示。

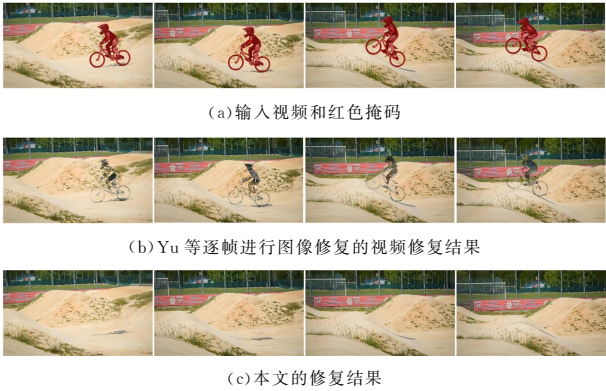


图 1 逐帧修复和直接进行视频修复的结果

Fig. 1 Result of frame-by-frame inpainting and direct video inpainting

直接对整个视频进行修复的方法可分为非学习的方法和学习的方 法。非学习的方法主要是基于块(patch)的方法^[17-22],通过从视频其他帧中找到相似的块来填充缺失区域。这类搜索算法通常具有较高的计算复杂性且缺乏对高级语义的理解,对于复杂的对象或掩码(mask),可能找不到缺失的区域。基于学习的方法主要是基于深度学习对视频进行修复^[23-32]。Xu 等^[23]提出了一种光流导向视频修复方法,使用合成的光流场来指导像素的传播,以填充视频中的孔区域。该方法可能存在在参考帧中找不到孔区域的问题。Lee 等^[24]提出了用于视频修复的复制和粘贴网络,该框架复制了视频其他帧中的附加信息。Lai 等^[25]提出了一种基于深度递归网络的方法,强制实施视频中的时间一致性。Chang 等^[26]提出了一种通过结合光流扭曲和基于图像的修复模型,该模型以时空一致的方式实现视频对象移除任务。CombCN^[27]通过 3D 卷积网络来实现时间一致性,通过 2D 卷积合成网络来提高视频质量,但仅在对齐的面部和具有矩形掩码的车辆等低分辨率视频上得到了验证,用于自由格式掩码和实际物体移除的情况有待验证。VINet^[28]引入了一种基于 3D-2D 的编码器-解码器模型,旨在收集和改进来自近邻帧的潜在内容,由于网络设计的特点,其时间窗口仅限于附近的帧,因此很难用在时间窗口外部可见的真实内容来修复孔区域。

目前仅有少量应用注意力机制进行图像修复的研究。Yu 等^[13]提出了一个上下文注意力层,该层搜索与粗略预测具有最高相似性的背景块集合。Yan 等^[16]介绍了一种由移位操作和导向损失驱动的 Shift-Net,移位操作推测编码器层中的上下文区域与解码器层中的关联孔区域之间的关系。将注意力机制应用于视频修复的研究目前仅有文献^[29],该文献提出了一种非对称注意块,以非局部的方式计算目标中孔边界像素与参考图像中非孔像素之间的相似点,对孔区域进行逐步修复。但是,非对称注意块存在复杂的迭代修复。这些方法可以利用相邻帧的信息,但忽略了语义相关性和特征连续性,存在语义错误和伪影的问题,难以保证局部像素的连续性。

针对上述问题,本文提出了连贯语义时空注意力层,该模型关注目标帧被遮挡而相邻帧可见的信息,以获取可见的内

容来填充目标帧的孔区域。CSSTA 层不仅可以对孔特征之间的语义相关性进行建模,而且还能对远距离信息和孔区域之间的远程关联进行建模,并在此基础上考虑到孔区域的空间一致性,对孔区域像素之间的语义相关性进行建模。有了注意力层,网络就可拥有一个无限的时空窗口,以参照相邻帧可用信息,用于和修复具有全局和局部语义一致内容的孔区域。本文提出的基于注意力的方法无需进行光流计算,因此能够处理具有遮挡或很难用光流建模的大孔洞的挑战性场景。将视频修复分为两个阶段,两个阶段网络都是编码器-解码器结构。第一个阶段训练一个粗糙的网络来粗略估计缺失区域的内容;第二个阶段嵌入了 CSSTA 层,并以第一阶段的输出为导向对缺失区域内容进行精细化修复。为了获得更好的修复结果,以前的方法使用 VGG Loss 来计算特征层次的 L_1 损失,但是 VGG^[33]主要面向的任务是图像分类,而针对图像修复/视频修复,最好的方式是利用图像到图像的深度学习对真值图像提取语义特征。本文提出了一种新的损失函数特征损失,训练了一个以原始真值图像为输入,以真值图像为输出的卷积自编码器网络^[34],利用自编码器网络提取特征,然后计算其与视频修复网络解码器的高级语义特征之间的距离。

2 时空注意力修复网络

本文采用的视频修复网络如图 2 所示。给定每一帧要填充的区域,目标是通过查看视频中的其他帧来填充空白区域。把要填充的图像称为目标帧,视频中的其他帧称为参考帧,算法用带有缺失区域的视频作为输入,包括帧 $\{I_t|t=1,\cdots,n\}$ 和每帧的目标修复掩码(mask) $\{M_t|t=1,\cdots,n\}$ 。当输出修复对象移动并且视点改变时,帧中被遮挡或删除的部分通常会在过去/将来的帧中显示出来,如果在时间范围内存在这样的提示,则可以借用那些可见的信息来恢复当前帧。

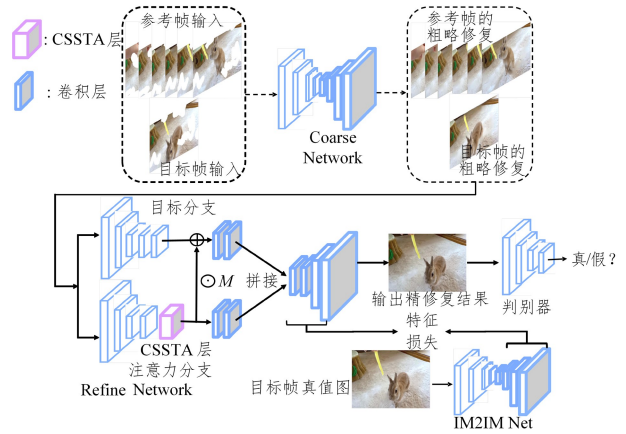


图 2 整体的视频修复网络结构框架

Fig. 2 Overall video inpainting network structure framework

2.1 连贯语义时空注意力

在卷积神经网络中,每个卷积核的尺寸都是有限的,只能处理局部内的信息。在单帧图像上,CNN 难以对距离较远的像素进行建模。在时间维度上,3D CNN 可以捕获时间信息,但计算量大且难以训练。 M 表示当前帧被遮挡或者不可见的信息, $\bar{M}$ 表示参考帧可见的信息,仅仅对 M 与 $\bar{M}$ 之间的关系进行建模是不够的,因为其忽略了孔区域的关系。

为了达到更好的视频修复效果,本文研究了人类在视频修复中的行为。人在视频修复的过程中首先观察单帧图像的整体结构,从相邻帧查看是否存在缺失部分的内容,用于构思当前帧缺失区域内容,从而保持视频时间的一致性。然后将相邻帧可见内容填充到当前帧,并构思相邻帧没有观察到的缺失内容。构思的内容要使得当前帧的孔区域内容保持空间结构一致性,这实际上保证了最终结果的局部像素连续性。

基于以上过程,提出了 CSSTA 层,不仅可以对孔特征之间的语义相关性进行建模,而且可以关注当前帧缺失而其他帧存在的信息来对当前帧缺失区域进行修复。网络有一个无限的时空窗口大小,用于参照无限个相邻帧的信息,因此可以对远距离信息和孔区域之间的远程关联进行建模。CSSTA 层分为帧间注意力和孔区域注意力。

(1) 帧间注意力。以运动中的物体为例,如图 3 中的箭头部分,首先从参考帧 $\bar{M}$ 中抽取块(patch)作为卷积核对目标 feature map 进行卷积,可以得到一个值向量,表示每个 patch 之间的互相关。图 3 中两个特征图中间位置 3×3 的块 m_i 和 $\bar{m}_i$ 表示目标特征图缺失的孔区域在参考特征图中能够找到,当以这部分块为卷积核进行卷积时,可以得到较大的值向量,获取到参考帧中可用的信息以对缺失信息进行补充,具体如式(1)所示:

$$S_{m_i \bar{m}_i} = \langle \frac{m_i}{|m_i|}, \frac{\bar{m}_i}{|\bar{m}_i|} \rangle \quad (1)$$

其中, m_i 表示目标帧特征图块, $\bar{m}_i$ 表示参考帧特征图块。

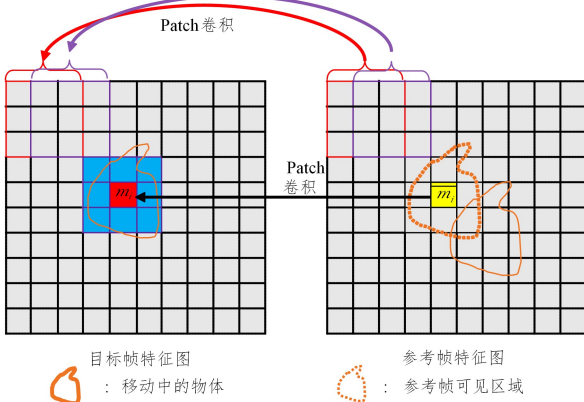


图 3 CSSTA 层帧间注意力

Fig. 3 Inter-frame attention of the CSSTA layer

(2) 孔区域注意力。为了对孔区域内的语义关系进行建模,如图 4 所示,蓝色部分表示缺失的孔区域,红色的 1×1 的块表示孔区域抽取的块 m_j ,使用块 m_j 作为 1×1 卷积核对目标帧孔区域进行卷积,以表示孔区域每个块之间的互相关。

$$S_{m_j m_j} = \langle \frac{m_j}{|m_j|}, \frac{m_j}{|m_j|} \rangle \quad (2)$$

其中, m_j 表示从孔区域抽取的作为卷积核的 1×1 的块, m_i 表示孔区域的 1×1 的块。实践中,抽取目标帧孔区域的 1×1 的块对目标帧的所有区域进行卷积,经过 softmax 得到注意力得分 $S_{m_j m_j}$,掩码区域的注意力得分则为 $S_{m_j m_j} \odot M_i$, $\odot$ 为像素级点积, M_i 为目标帧掩码,注意力层最后的输出为帧间注意力和孔区域注意力相加,注意力层的输出表示为:

$$S_{att} = S_{m_i \bar{m}_i} + \overline{S_{m_j m_j}} \odot M_i \quad (3)$$

其中, $S_{m_i \bar{m}_i}$ 为帧间注意力输出, $\overline{S_{m_j m_j}} \odot M_i$ 为孔区域注意力输出,式(3)用于计算图 2 中注意力层的输出。

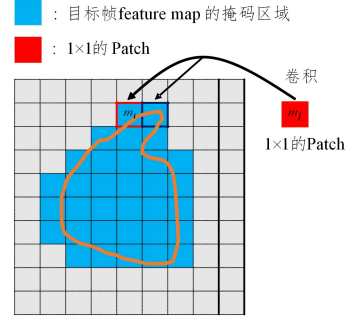


图 4 CSSTA 层孔区域注意力(电子版为彩色)

Fig. 4 Hole region attention of the CSSTA layer

2.2 Feature Loss

VGG 损失首先在文献[35]中被提出,用于保留图像内容进行风格转换,目前广泛应用于图像修复^[10,36]和超分辨率^[37-38]中,以减轻由于 L_1 损失而造成的模糊。然而,VGG 主要用于图像分类任务,而针对图像修复/视频修复,最好的方式是利用图像到图像的深度网络对真值(ground truth)图像提取高级语义特征。图像分类任务中,如对猫的图像进行分类,分类模型主要学习到的是图片中猫的语义特征,而对图片中其他部分的语义特征的表达不足。在修复过程中,孔区域的位置是随机的,因此模型需要学习到图像全局的语义信息。而输入为真值图像,输出等于真值图像的深度网络能够学习到图像全局的语义信息。

基于以上原理,本文提出了一种新的损失函数特征损失 (Feature Loss),训练一个以原始真值图像为输入,以真值图像为输出的卷积自编码器网络(IM2IM Net)。IM2IM Net 能够学习到图像的高级语义特征,利用这个自编码器网络提取语义特征,然后计算与视频修复网络解码器的高级语义特征之间的距离。视频修复网络和 IM2IM Net 均为编码器-解码器结构,视频修复网络编码器阶段学习到了足够多的特征,解码器阶段逐步恢复图像,解码器阶段应包含原始图像应有的特征。视频修复网络解码器应与 IM2IM Net 的解码器有相似的语义信息,IM2IM Net 在 Places2^[39]数据集上进行训练,将原始图像作为输入,输出为原始图像,计算输出与原始图像的 L_1 损失,并利用 GAN 进行联合训练,IM2IM Net 很容易收敛,几乎能生成与原始图像一样的图像,Feature Loss 在特征层计算 L_1 损失。

$$\mathcal{L}_F = \sum_{i=1}^n \sum_{p=0}^{P-1} \frac{|\Psi_p^{Y_i} - \Psi_p^{\hat{Y}_i}|}{N_{\Psi_p^{Y_i}}} \quad (4)$$

其中, $\Psi_p^{Y_i}$ 表示在给定输入 Y_i 的情况下 IM2IM 网络的解码器第 p 层激活输出得到的高级语义特征,而 $N_{\Psi_p^{Y_i}}$ 是第 p 层中的元素数量。使用 IM2IM Net 的解码器的后面 3 层来计算 Feature Loss。Feature Loss 有效的原因是 IM2IM Net 的解码器和视频修复模型的解码器都是在逐步恢复视频的原始信息,而两者的解码器结构一样,那么 IM2IM Net 和视频修复模型解码器的每一层的输出应该有相近的高级特征信息,而计算两个解码器的高级语义特征之间的 L_1 损失,使得视频修

复模型解码器能够输出与 IM2IM Net 解码器相近的高级语义信息。

2.3 谱归一化判别器

以前的图像修复网络始终使用局部判别器来改善结果,但是局部判别器不适用于不规则的孔,该孔可以是任何形状和位于任何位置。受 Gated Conv^[14] 启发,本文引用了谱归一化特征判别器,通过检查图像修复网络的特征图来区分修复的图像和原始图像,使用谱归一化来进一步稳定 GANs 的训练。另外,本文使用(hinge loss)铰链损失作为目标函数来判别输入的真假:

$$\mathcal{L}_D = \mathbb{E}_{x \sim P_{\text{data}}(x)} [\text{ReLU}(1 + D(x))] + \mathbb{E}_{z \sim P_z(z)} [\text{ReLU}(1 - D(G(z)))] \quad (5)$$

$$\mathcal{L}_G = -\mathbb{E}_{z \sim P_z(z)} [D(G(z))] \quad (6)$$

其中, G 是视频修复网络,输入为视频 z , D 是判别器。

2.4 网络结构

整体的网络是一个双阶段编码器-解码器结构,编码器的输入包括 RGB 图像和空白掩码。这些输入沿通道轴连接,然后被送入第一层。使用门控卷积层作为基本的构建块,因为它对处理孔之类的无效信息非常有效。编码器下采样的特征映射到原来的 $1/4$ 规模,以确保高频细节,利用膨胀(dilation)卷积进一步放大大接受域的大小。如图 2 所示,总的输入为一个视频帧序列 $\{I_t | t = 1, \dots, n\}$ 和掩码序列 $\{M_t | t = 1, \dots, n\}$,经过第一阶段网络得到基于单帧的图像修复序列 $\{P_t | t = 1, \dots, n\}$,第二阶段是一个双分支网络,上面一个分支为目标分支(target branch),下面一个分支为注意力分支(attention branch),参考帧和目标帧输入到下面的注意力分支,目标帧输入到目标分支,注意力分支经过注意力层得到注意力输出 $S_{m_i, \overline{m_i}}$ 。为了增强注意力的融合,把注意力输出的掩码部分与目标分支的中间结果 F_t 融合为 $F_t^{\text{att}} = F_t + S_{m_i, \overline{m_i}} \odot M_t$,利用注意力分支的输出来重建目标图像的空白区域,让解码器专注于通过损失函数来恢复空白区域,而损失函数对该区域施加了更多的权重,注意力分支和目标分支的输出按维度拼接(concat)起来馈送到解码器,从而得到最终的输出。

整个视频修复网络会对视频逐帧进行修复,为了解决输出视频经常显示闪烁伪像的问题,在视频修复的解码器后面附加一个时间一致性网络,对视频进行后处理。受 Lai 等^[25] 的启发,为了达到前一帧的时间稳定性和当前帧的感知相似性之间的平衡,在训练时配备有卷积 GRU 的编码器-解码器网络,对视频修复网络进行训练收敛后,把视频修复的输出作为时间一致性网络的输入,输出为时间一致的修复视频。

2.5 总体损失函数

整体网络结构的损失为:

$$\mathcal{L} = \lambda_R \mathcal{L}_R + \lambda_F \mathcal{L}_F + \lambda_G \mathcal{L}_G \quad (7)$$

其中, $\mathcal{L}_R$ 是重建损失, $\mathcal{L}_F$ 是 Feature Loss, $\mathcal{L}_G$ 是生成器损失,3 个损失的平衡权重 $\lambda_R, \lambda_F, \lambda_G$ 通过实验分别设置为 1, 0.5, 1。 $\mathcal{L}_R$ 包含掩码区域和非掩码区域的 $\mathcal{L}_1$ 损失。

$$\mathcal{L}_1 = \delta M_t \odot \|\hat{Y}_t - Y_t\|_1 + (1 - M_t) \odot \|\hat{Y}_t - Y_t\|_1 \quad (8)$$

其中, $\hat{Y}_t, Y_t$ 表示网络的输出和未破损的图片, $\mathcal{L}_1$ 是像素级别的 $\mathcal{L}_1$ 损失, δ 是一个值为 100 的常数,目的是增加重建区域的损失权重,使得模型更加关注重建区域。

3 实验

3.1 数据准备与评价指标

本文考虑了两种常见的修复场景。第一种是删除不必要的前景对象,给出一个掩码(mask)来指示前景目标对象的区域。在第二种场景中,目的是修复视频中自由格式的任意区域,其中包含前景或背景。这种场景对应于一些实际应用程序,如水印去除和视频恢复。由于视频修复不需要标注数据,只需要生成自由格式的视频掩码,将视频掩码区域置零并破坏掉就可以得到需要修复的视频。而未被破坏掉的视频则为真值图像。网络以破损的视频和导向掩码为输入,得到修复后的视频。为了增加视频的复杂场景以及多样性,使用 YouTube-VOS^[40] 数据集进行训练,使用自由格式掩码生成算法生成自由格式的掩码。视频修复也常常用于视频目标对象移除。为了模拟这种场景,给出了一个掩码(mask)来勾勒前景对象的区域,使用 DAVIS^[41] 数据集提供像素级对象掩码,并将其覆盖到视频的每一帧上,mask 帧和非 mask 帧构成训练对。最后使用 YouTube-VOS 的 4453 个视频进行训练。使用 YouTube-VOS 的 507 个视频和 DAVIS 的 150 个视频进行测试。

本文用峰值信噪比(Peak Signal to Noise Ratio, PSNR)和结构相似性(Structural Similarity, SSIM)测量修复后的视频相比原始视频的修复程度,PSNR 测量图像的失真,SSIM 测量两幅图像在结构上的相似性。

3.2 实验设置

本文算法使用 Pytorch v1.2 实现,使用 256×256 的图像进行训练和定量比较,在训练过程中采用了诸如翻转等数据扩充;使用 DAVIS 数据的原始图像进行定性比较,提高图像显示效果;从视频中每隔 10 帧选取一帧进行显示,以明显区分各视频帧。训练过程中,参考帧数设置为 6,对于所有的实验,mini-batch 大小设置为 1,每次训练一个视频,优化器使用 Adam,初始化学率设置为 1×10^{-4} ,每迭代 10 次,学习率乘以 0.9,整个过程使用 4 块 NVIDIA GTX 1080Ti GPUs 训练 4 天。

3.3 定性比较

本文方法和现有的基于学习的 3 种方法进行了比较,在本文的测试视频和掩码上,运行它们的测试代码和本文的测试代码。

(1) Yu 等提出的 Gated Conv。基于前馈卷积网络的单帧图像修复算法,对视频进行逐帧修复^[14]。

(2) Xu 等提出的 DFGNet。光流导向的视频修复算法,通过估计缺失部分的光流场来达到修复视频的目的^[22]。

(3) Kim 等提出的 VINet。其引入了一种新颖的 3D-2D 前馈网络,该网络建立在基于 2D(基于图像)的编码器-解码器模型之上,旨在收集和改进来自参考帧的潜在内容^[28]。

为了实验的公平性,在自由格式的视频掩码和真实视频目标掩码上进行测试。对于视频目标移除,使用 DAVIS 的视频,其中每个视频帧都有一个对象的像素级注释。如图 5 所示, Yu 等的结果出现了扭曲的结构和颜色混淆,并且在生成的图像上仍然可以观察到一些伪像,这主要是由于对视频进行逐帧修复,没有考虑到时间维度的视频帧之间的关联性。Xu 等的

结果出现了不合理的黑线纹理和伪影,在某些情况下,缺失区域不能被光流跟踪的已知像素填充(例如,当前需要修复的目标帧的缺失区域在参考帧中找不到缺失的信息),这意味着模型未能将某些缺失区域连接到任何像素。Kim 等提出的模型的性能较好,但其预测在一定程度上是模糊和缺乏细节的,并且伴有一定的伪影。本文的注意力机制在重点关注参考帧可用信息的同时,也对生成内容之间的关系进行了建模,获得了良好的修复结果。图 6 给出了 DAVIS 数据集的一些视频目标移除的案例。



图 5 可视化视频目标移除的定性对比

Fig. 5 Qualitative comparison of visual video object removal

对于自由格式的视频掩码修复,使用 YouTube-VOS 的视频,然后随机生成自由格式的对应掩码,形成测试对,对比的 3 种方法使用同一个视频和对应的掩码,视频序列帧修复结果如图 7 所示。

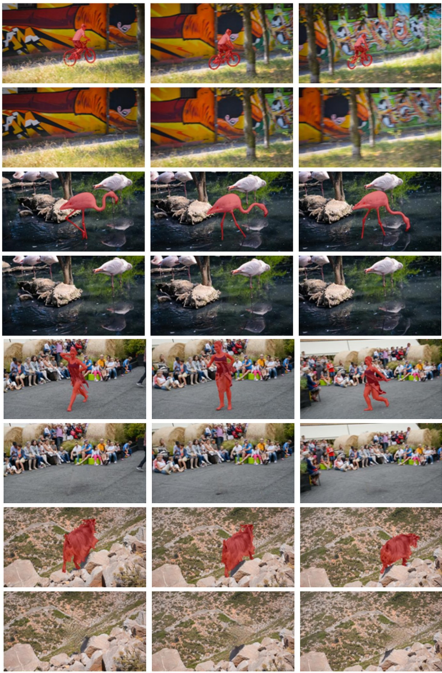


图 6 视频目标移除案例

Fig. 6 Video object removal case



图 7 可视化自由格式掩码的定性对比

Fig. 7 Qualitative comparison of visualized free-form mask

3.4 定量比较

本文利用来自 YouTube-VOS 的 507 个视频和 DAVIS 的 150 个视频进行测试。对于物体移除的情况,无法获得真实的标签,在已有的视频基础上合成虚拟物体,即使用 DAVIS 的视频对象掩码,打乱 DAVIS 的视频顺序,将视频掩码复制到视频序列,形成虚拟的物体,并用于比较 PSNR 和 SSIM。在表 1 和表 2 中,将本文方法与基于单帧图像修复^[14]和视频修复^[22,28]的方法进行了定量比较,表 1 和表 2 中的执行时间表示平均每个视频的修复时间,平均每个视频有 71.54 帧。可以看出,本文方法在自由格式掩码和目标掩码上均优于其他方法,并且几乎实时运行。

表 1 自由格式掩码定量评价

Table 1 Quantitative evaluation of free-form mask

	YouTube-VOS			DAVIS		
	PSNR	SSIM	Time/s	PSNR	SSIM	Time/s
Yu 等的方法	24.418	0.788	1.40	23.623	0.757	1.40
Xu 等的方法	24.04	0.858	934	25.16	0.824	934
Kim 等的方法	21.207	0.748	15.8	21.619	0.750	15.8
本文方法	27.939	0.906	12.3	28.651	0.815	12.3

表 2 对象掩码定量评价

Table 2 Quantitative evaluation of object mask

	YouTube-VOS			DAVIS		
	PSNR	SSIM	Time/s	PSNR	SSIM	Time/s
Yu 等的方法	24.537	0.822	1.40	23.913	0.797	1.40
Xu 等的方法	25.77	0.890	934	26.04	0.883	934
Kim 等的方法	25.775	0.875	15.8	25.325	0.864	15.8
本文方法	28.127	0.926	12.3	28.744	0.956	12.3

3.5 消融研究

(1)CSSTA 层的影响。为了研究 CSSTA 层的影响,本文将 CSSTA 层从模型中删除,然后利用同样的数据和训练方法对没有 CSSTA 层的模型进行训练,并在 YouTube-VOS 的

507 个测试视频上进行自由格式的掩码测试。对比结果如表 3 所列, 添加了 CSSTA 层后, PSNR 和 SSIM 有明显的提升。在 DAVIS 数据集上进行对象掩码测试, 结果如图 8 所示。没有 CSSTA 层时, 孔区域中心呈现出扭曲的结构, 这可能是模型没有获取到参考帧足够多的信息, 对于所有的信息做同样的处理; 添加了 CSSTA 以后, 重点关注当前帧缺失而参考帧存在的信息, 取得了更加理想的修复效果。

表 3 CSSTA 层消融研究
Table 3 CSSTA layer ablation study

	PSNR	SSIM
No CSSTA	18.227	0.721
CSSTA	27.939	0.906



图 8 CSSTA 层的影响

Fig. 8 Impact of CSSTA layer

(2) Feature Loss 的影响。为了研究 Feature Loss 的影响, 本文将 Feature Loss 从损失函数中删除, 然后利用同样的数据和训练方法对没有 Feature Loss 的模型进行训练, 并在 YouTube-VOS 的 507 个测试视频上进行自由格式的掩码测试。对比结果如表 4 所列, 增加 Feature Loss 后 PSNR 和 SSIM 均有提升。在 DAVIS 数据集上进行了对象掩码测试, 如图 9 所示。在没有 Feature Loss 的情况下, 修复结果的孔区域呈现出模糊的情况, 这可能是由于训练不稳定和对图像语义的误解造成的, Feature Loss 用于计算高级语义特征的损失, 有助于避免这些问题。

表 4 Feature Loss 消融研究
Table 4 Feature Loss ablation study

	PSNR	SSIM
No FL	25.07	0.841
FL	27.939	0.906



图 9 Feature Loss 的影响

Fig. 9 Impact of Feature Loss

结束语 本文提出了一种新的基于深度生成模型的视频修复模型, 设计了一种新的连贯语义时空注意力层, 可以同时目标帧孔区域和参考帧可见区域之间的关系以及孔之间的语义相关性进行建模, 引入了 Feature Loss 增强 CSSTA 层的学习能力, 提高训练的稳定性以实现更好的修复结果。实验结果表明, 相比其他方法, 本文方法不但加快了收敛速度, 还减少了迭代次数, 同时取得了具有竞争力的视频修复结果, 而且几乎实时运行, 从而证明本文提出的连贯语义时空注意

力网络在视频修复领域中的可行性。下一步将针对训练过程中速度较慢、占用内存大以及生成式网络具有不稳定性等问题进行深入研究。

参 考 文 献

[1] BARNES C, SHECHTMAN E, FINKELSTEIN A, et al. Patch-Match: A randomized correspondence algorithm for structural image editing[J]. ACM Transactions on Graphics, 2009, 28(3): 24.

[2] HE K, SUN J. Image completion approaches using the statistics of similar patches[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(12): 2423-2435.

[3] HO L J, CHOI I, KIM M H. Laplacian patch-based image synthesis[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, IEEE Computer Society, 2016: 2727-2735.

[4] KOMODAKIS N. Image completion using global optimization [C]// 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York City: IEEE, 2006: 442-452.

[5] LE MEUR O, EBDELLI M, GUILLEMOT C. Hierarchical super-resolution-based inpainting[J]. IEEE Transactions on Image Processing, 2013, 22(10): 3779-3790.

[6] XU Z, SUN J. Image inpainting by patch propagation using patch sparsity [J]. IEEE Transactions on Image Processing, 2010, 19(5): 1153-1165.

[7] SHAO X W, LIU Z K, SONG B. An Adaptive Image Inpainting Approach Based on TV Model[J]. Journal of Circuits and Systems, 2004(2): 113-117.

[8] LI Z D, HE H J, YIN Z K, et al. Adaptive Image Inpainting Algorithm Based on Patch Structure Sparsity[J]. Acta Electronica Sinica, 2013, 41(3): 549-554.

[9] IIZUKA S, SIMO-SERRA E, ISHIKAWA H. Globally and locally consistent image completion [J]. ACM Transactions on Graphics (ToG), 2017, 36(4): 1-14.

[10] LIU G, REDA F A, SHIH K J, et al. Image inpainting for irregular holes using partial convolutions[C]// Proceedings of the European Conference on Computer Vision (ECCV). Munich: Springer International Publishing, 2018: 85-100.

[11] PATHAK D, KRAHENBUHL P, DONAHUE J, et al. Context encoders: Feature learning by inpainting[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, IEEE Computer Society, 2016: 2536-2544.

[12] YANG C, LU X, LIN Z, et al. High-resolution image inpainting using multi-scale neural patch synthesis[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE Computer Society, 2017: 6721-6729.

[13] YU J, LIN Z, YANG J, et al. Generative image inpainting with contextual attention[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 5505-5514.

[14] YU J, LIN Z, YANG J, et al. Free-form image inpainting with gated convolution[C]// Proceedings of the IEEE International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019: 4471-4480.

- [15] SONG Y, YANG C, LIN Z, et al. Contextual-based image inpainting: Infer, match, and translate[C]// Proceedings of the European Conference on Computer Vision (ECCV). Munich: Springer International Publishing, 2018: 3-19.
- [16] YAN Z, LI X, LI M, et al. Shift-net: Image inpainting via deep feature rearrangement[C]// Proceedings of the European Conference on Computer Vision (ECCV). Munich: Springer International Publishing, 2018: 1-17.
- [17] HUANG J B, KANG S B, AHUJA N, et al. Temporally coherent completion of dynamic video[J]. ACM Transactions on Graphics (TOG), 2016, 35(6): 1-11.
- [18] NEWSON A, ALMANSA A, FRADET M, et al. Video inpainting of complex scenes[J]. Siam Journal on Imaging Sciences, 2014, 7(4): 1993-2019.
- [19] BERTALMIO M, BERTOZZI A L, SAPIRO G. Navier-stokes, fluid dynamics, and image and video inpainting[C]// Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Kauai, HI, USA: IEEE Computer Society, 2001.
- [20] PATWARDHAN K A, SAPIRO G, BERTALMÍO M. Video inpainting under constrained camera motion[J]. IEEE Transactions on Image Processing, 2007, 16(2): 545-553.
- [21] GAN L, GUO X Y, LIU B. A block matching algorithm for video inpainting[J]. Computer Applications and Software, 2013(9): 64-66.
- [22] XIAO C X, LIU S, LIN C C, et al. A Global Space-Time Optimization Framework for Video Completion[J]. Journal of Computer-Aided Design & Computer Graphics, 2008, 20(9): 1204-1211.
- [23] XU R, LI X, ZHOU B, et al. Deep flow-guided video inpainting[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019: 3723-3732.
- [24] LEE S, OH S W, WON D Y, et al. Copy-and-paste networks for deep video inpainting[C]// Proceedings of the IEEE International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019: 4413-4421.
- [25] LAI W S, HUANG J B, WANG O, et al. Learning blind video temporal consistency[C]// Proceedings of the European Conference on Computer Vision (ECCV). Munich: Springer International Publishing, 2018: 170-185.
- [26] CHANG Y L, LIU Z Y, HSU W, et al. Vornet: Spatio-temporally consistent video inpainting for object removal[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, CA, USA: IEEE, 2019: 1785-1794.
- [27] WANG C, HUANG H, HAN X, et al. Video inpainting by jointly learning temporal structure and spatial details[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 5232-5239.
- [28] KIM D, WOO S, LEE J Y, et al. Deep video inpainting[C]// proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019: 5792-5801.
- [29] OH S W, LEE S, LEE J Y, et al. Onion-peel networks for deep video completion[C]// Proceedings of the IEEE International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019: 4403-4412.
- [30] CHANG Y L, LIU Z Y, LEE K Y, et al. Learnable gated temporal shift module for deep video inpainting[J]. arXiv: 1907.01131, 2019.
- [31] KIM D, WOO S, LEE J Y, et al. Deep blind video decaptioning by temporal aggregation and recurrence[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019: 4263-4272.
- [32] CHANG Y L, LIU Z Y, LEE K Y, et al. Free-form video inpainting with 3d gated convolution and temporal patchgan[C]// Proceedings of the IEEE International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019: 9066-9075.
- [33] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv: 1409.1556, 2014.
- [34] KINGMA D P, WELING M. Auto-encoding variational bayes[J]. arXiv: 1312.6114, 2013.
- [35] GATYS L A, ECKER A S, BETHGE M. A neural algorithm of artistic style[J]. arXiv: 1508.06576, 2015.
- [36] NAZERI K, NG E, JOSEPH T, et al. Edgeconnect: Generative image inpainting with adversarial edge learning[J]. arXiv: 1901.00212, 2019.
- [37] JOHNSON J, ALAHI A, LI F F. Perceptual losses for real-time style transfer and super-resolution[C]// European Conference on Computer Vision. Amsterdam: Springer International Publishing, 2016: 694-711.
- [38] LEDIG C, THEIS L, HUSZÁR F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE Computer Society, 2017: 4681-4690.
- [39] ZHOU B, LAPEDRIZA A, KHOSLA A, et al. Places: A 10 million image database for scene recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(6): 1452-1464.
- [40] XU N, YANG L, FAN Y, et al. Youtube-vos: A large-scale video object segmentation benchmark[J]. arXiv: 1809.03327, 2018.
- [41] PERAZZI F, PONT-TUSET J, MCWILLIAMS B, et al. A benchmark dataset and evaluation methodology for video object segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE Computer Society, 2016: 724-732.



LIU Lang, born in 1994, postgraduate. His main research interests include deep learning and pattern recognition.



DAN Yuan-hong, born in 1981, Ph.D., associate professor. His main research interests include pattern recognition and intelligent control.