

基于残差卷积注意力网络的视频修复

李德财, 严群, 姚剑敏, 林志贤, 董泽宇

引用本文:

李德财, 严群, 姚剑敏, 等. 基于残差卷积注意力网络的视频修复[J]. 液晶与显示, 2022, 37(1): 86–96.

LI De-cai, YAN Qun, YAO Jian-min, et al. Video inpainting based on residual convolution attention network[J]. *Chinese Journal of Liquid Crystals and Displays*, 2022, 37(1): 86–96.

在线阅读 View online: <https://doi.org/10.37188/CJLCD.2021-0196>

您可能感兴趣的其他文章

Articles you may be interested in

基于生成对抗网络的图像超分辨率重建算法

Image super resolution reconstruction algorithm based on generative countermeasure network

液晶与显示. 2021, 36(12): 1720–1727 <https://doi.org/10.37188/CJLCD.2021-0227>

时空域深度学习火灾烟雾检测

Spatio-temporal deep learning fire smoke detection

液晶与显示. 2021, 36(8): 1186–1195 <https://doi.org/10.37188/CJLCD.2020-0239>

基于改进Faster R-CNN的交通标志检测算法

Traffic sign detection algorithm based on improved Faster R-CNN

液晶与显示. 2021, 36(3): 484–492 <https://doi.org/10.37188/CJLCD.2020-0195>

融合注意力机制的移动端人像分割网络

Mobile-based portrait segmentation network with attention mechanism

液晶与显示. 2020, 35(6): 547–554 <https://doi.org/10.3788/YJYXS20203506.0547>

基于局部期望最大化注意力的图像降噪

Local expectation-maximization attention network for image denoising

液晶与显示. 2020, 35(4): 350–359 <https://doi.org/10.3788/YJYXS20203504.0350>

文章编号:1007-2780(2022)01-0086-11

基于残差卷积注意力网络的视频修复

李德财¹, 严 群^{1,2*}, 姚剑敏^{1,2}, 林志贤¹, 董泽宇¹

(1. 福州大学 物理与信息工程学院, 福建 福州 350108;

2. 晋江市博感电子科技有限公司, 福建 晋江 362200)

摘要: 视频修复旨在填补视频中的缺失区域, 由于很难精确保持修复内容的时空一致性, 故视频修复仍具有挑战性。针对现有视频修复中存在的修复结果语义信息不连续, 出现视频模糊和时间伪影, 以及网络设计越来越复杂, 网络整体速度变慢的问题, 本文提出了一种基于残差网络的卷积注意力网络(RCAN)用以视频修复。通过将自注意力机制和全局注意力机制引入进残差网络, 增强网络对所有输入帧的时空特征的学习能力, 并采用时空对抗损失函数进行优化, 提高视频修复的质量。同时网络还能够高度自由地定义层数和参数量, 提高网络的实际应用能力。实验结果表明, 该网络在 DAVIS 和 YouTube-VOS 数据集上取得了 PSNR 为 30.68 dB, SSIM 为 0.961, FID 为 0.113 的平均修复结果, 基本符合实际场景对模型的修复质量要求, 为视频修复提供了一种新思路。

关键词: 深度学习; 视频修复; 自注意力机制; 残差网络; 生成对抗网络

中图分类号: TP391.41; TH691.9 **文献标识码:** A **doi:** 10.37188/CJLCD.2021-0196

Video inpainting based on residual convolution attention network

LI De-cai¹, YAN Qun^{1,2*}, YAO Jian-min^{1,2}, LIN Zhi-xian¹, DONG Ze-yu¹(1. College of Physics and Information Engineering, Fuzhou University,
Fuzhou 350108, China;2. Jinjiang RichSence Electronic Technology Company Limited,
Jinjiang 362200, China)

Abstract: Video inpainting, which aims at filling in missing regions of a video, remains challenging due to the difficulty of preserving the precise spatial and temporal coherence of video contents. In order to solve the problems of discontinuous semantic information, video blurriness and temporal artifact, and more and more complex network design, the overall speed of the network becoming slow, this paper proposes a residual convolution attention network (RCAN) for video inpainting. By introducing the self-attention mechanism and the global attention mechanism into the residual network,

收稿日期: 2021-07-24; **修订日期:** 2021-09-26.

基金项目: 国家重点研发计划(No.2016YFB0401503); 广东省科技重大专项(No.2016B090906001); 福建省科技重大专项(No.2014HZ003-1); 广东省光信息材料与技术重点实验室开放基金(No.2017B030301007)

Supported by National Key Research and Development Program of China(No. 2016YFB0401503); Science and Technology Major Program of Guangdong Province (No.2016B090906001); Science and Technology Major Program of Fujian Province (No.2014HZ003-1); Guangdong Provincial Key Laboratory of Optical Information Materials and Technology (No.2017B030301007)

* 通信联系人, E-mail: qunfyan@gmail.com

the ability of the network to learn the spatio-temporal features of all input frames is enhanced. This method proposes a spatial-temporal adversarial loss function to optimize RCAN, which improves the quality of video inpainting. At the same time, the network can define the number of layers and parameters with a high degree of freedom to improve the practical application ability of the network. Experimental results show that the network can achieve an average inpainting result in that the PSNR is 30.68 dB, the SSIM is 0.961, and the FID is 0.113 on DAVIS and YouTube-VOS data sets. This method meets the inpainting quality requirements of the actual scene on the model and provides a new idea for video inpainting.

Key words: deep learning; video inpainting; self-attention mechanism; residual networks; generative adversarial networks

1 引言

视频修复旨在用视频中时空相关的内容来填充视频序列中的缺失区域^[1]。高质量的视频修复对损坏的视频修复^[2-3],未损坏的对象移除^[4-5],视频重新定向^[6]和曝光不足或过度的视频修复^[7]等任务具有重要意义。高质量的视频修复仍然面临着许多挑战,例如缺乏对视频的高度理解^[8]和高度的计算复杂性。大多数现有的视频修复算法^[9-14]遵循传统的图像修复方式,通过将问题公式化为基于块拼接的优化任务,通过采样空间或空间-时间匹配来解决已知区域的最小化问题。尽管有一些好的修复结果,但这些方法不能处理复杂运动的情况,且计算复杂度较高。

随着深度卷积神经网络的发展,深度视频修复取得了重大进展。其中使用3D卷积和递归网络进行视频修复^[15-16]越来越受到人们的关注。文献[6]采用循环网络来确保时间的一致性。文献[16]首次提出将3D和2D全卷积网络相结合来学习视频的时间信息和空间细节。Xu等人^[17-18]通过联合估计空间特征和光流来提高性能。Chang等人提出了用于自由形式视频修复的时态SN-PatchGAN^[19]和时态移位模块^[20]。于冰等人^[21]提出基于时空特征的生成对抗网络进行视频修复。这些方法通常会从附近的帧中收集信息来填充丢失的区域,但由于有限的时间感受野,在每一帧上直接应用图像修复算法会导致时间伪影和抖动。为解决上述挑战,目前最先进的方法是应用注意力模块来捕捉长距离帧的对应关系,从而可以使用来自远处帧的内容来填充目标帧中的缺失区域^[22-23]。文献[22]是通过将视频

帧进行加权求和来合成缺失的内容,并进行逐帧关注。文献[23]通过逐像素的关注,从边界向内逐渐填充具有相似像素的缺失区域。虽然有较好的修复结果,但由于视频中复杂的运动会引起明显的外观变化,导致修复帧的匹配效果较差,无法对复杂的空间信息进行合理匹配。此外由于没有对复杂的时间信息处理模块进行针对性优化,而是对所有视频进行逐帧处理,导致修复结果的时间一致性较差,而且需要较大的计算量和较长的处理时间。

针对上述问题,本文提出了一种基于残差网络^[24]的卷积注意力网络(RCAN)用以视频修复。该网络以视频所有帧作为输入,使用自注意力机制和全局注意力机制^[25]来提取所有帧的时空特征信息,进而对所有输入帧进行修复。此外残差网络的引入,能够增强深度网络的性能,提高网络对所有帧的时空特征的学习能力,同时网络采用时空对抗损失函数进行优化,提高模型泛化能力,实现时空一致的视频修复。此外网络还能够高度自由地定义层数和参数量,降低计算复杂度,减少计算资源需求,加快网络训练速度,大幅度提高实际应用能力。

2 残差卷积注意力网络

2.1 网络概述

视频修复的目标是用时空连贯的内容来填补视频序列中的缺失区域。一般来说,当前帧中的遮挡区域可能会在远距离帧的区域中显示出来,所以以相邻帧和远距离帧为基础,从整个视频中选取有用的内容,来修复目标帧中缺失的区域,是一种更有效的方法。设 $X_1^T = \{X_1, X_2, \dots, X_T\}$

是高为 H , 宽为 W , 帧数为 T 的损坏视频帧, $M_1^T = \{M_1, M_2, \dots, M_T\}$ 表示相应的逐帧掩膜, $Y_1^T = \{Y_1, Y_2, \dots, Y_T\}$ 是原始视频帧, $\hat{Y}_1^T = \{\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_T\}$ 是模型生成的视频帧。学习经过掩膜处理的损坏视频帧 X_1^T 到模型生成的视频帧 $\hat{Y}_1^T$ 的映射函数, 使得真实数据 $p(Y_1^T | X_1^T)$ 的条件分布可以由生成的数据 $p(\hat{Y}_1^T | X_1^T)$ 的条件分布来近似。基于马科夫假设^[26], 将问题简化为:

$$p(\hat{Y}_1^T | X_1^T) = \prod_{t=1}^T p(\hat{Y}_{t-n}^{t+n} | X_{t-n}^{t+n}, X_{1,s}^T), \quad (1)$$

其中: X_{t-n}^{t+n} 代表以 t 时刻为中心, 左右距离为 n 的相邻帧, $X_{1,s}^T$ 是在 X_1^T 的基础上, 以 s 速率均匀采样的远距离帧, 包含了整个视频的大部分关键帧, 可以有效描述视频内容, 从而避免视频场景切换带来的修复效果差的问题。视频修复不仅与视频相邻帧保持时空一致, 还需要与视频的远距离关键帧保持时空一致。

本文提出的残差卷积注意力网络, 以目标帧的相邻帧和远距离帧即视频所有帧作为输入, 对所有视频输入帧进行修复。通过将自注意力机制和全局注意力机制引入残差网络中, 增强网络对所有帧的空间及时间维度信息的学习能力, 保持与相邻帧以及远距离关键帧的时空一致, 提高

视频修复效果。

残差卷积注意力网络结构如图 1 所示, 由初始化处理模块、残差卷积注意力模块以及解码模块构成。其中初始化处理模块包含位置编码以及 3D 卷积模块, 残差卷积注意力模块由轻量型多头自注意力机制、残差前馈网络及残差感知单元组成, 解码模块由反卷积的叠加并添加残差连接构成。网络输入为 $f_1^T = \{f_1, f_2, \dots, f_T\}$ 的视频序列, 其中每个视频序列为 $f_i \in R^{h \times w \times c}$, 大小为 $384 \times 192 \times 3$, 之后通过初始化处理模块, 将输入视频帧下采样为 $1/4$ 大小, 再通过堆叠残差卷积注意力模块, 提取输入帧的时空特征, 其中输入帧特征比例依次变为 $1/8, 1/16, 1/32, 1/64$ 。然后进入解码模块, 通过上采样将特征反向解码, 最后经过激活函数得到修复结果。其中本文使用典型的通道数, 即 $C_1 = 32, C_2 = 64, C_3 = 128, C_4 = 256, C_5 = 512$ 。此外结构图中的 $\times L_i$ 为残差层数即 L-MHAS 和 RFFN 模块的重复次数, 本文采用的 L_i 为 $[4, 8, 8, 4]$ 。在实际应用中可以控制其残差层数来控制模型深度, 以此来控制模型的复杂度和计算量, 使得该网络既可以作为大的网络结构使用, 以追求更好的性能, 也可以作为一个小的轻量化网络进行使用, 减少计算量提高速度。

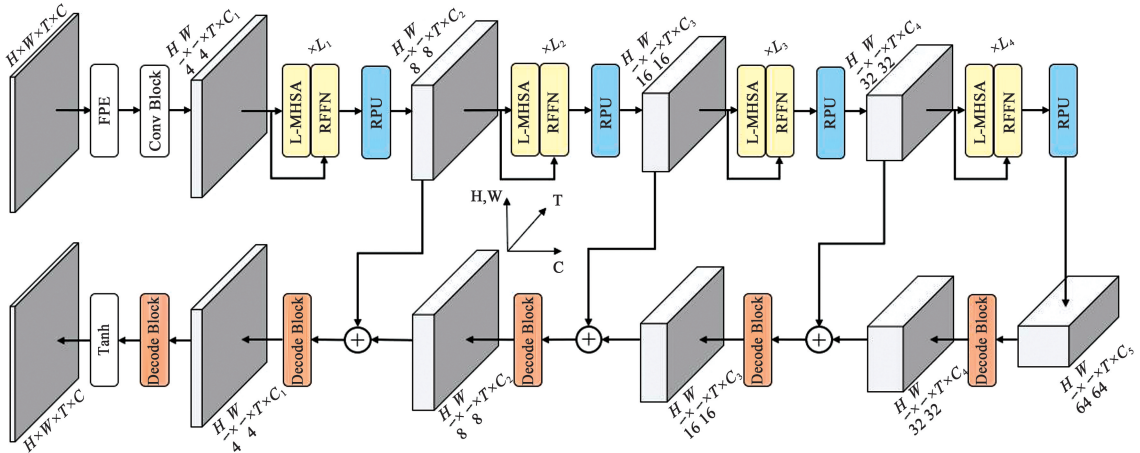


图 1 残差卷积注意力网络结构

Fig.1 Residual convolution attention network structure

2.2 残差卷积注意力模块

残差卷积注意力模块由轻量型多头自注意力机制, 残差前馈网络以及残差感知单元构成, 每部分具体描述如下。

2.2.1 轻量型多头自注意力机制

自注意力机制由自注意层构成, 在自注意层中, 输入向量通过 3 个不同的投影矩阵转换为 3 个不同的向量, 即查询向量 q 、键向量 k 与值向量

v , 各向量的维度相同。之后将不同输入获得的向量合为 3 个不同的矩阵 Q, K 和 V , 进行放缩点积计算, 具体以计算过程如下:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (2)$$

自注意层进一步完善, 形成了多头自注意力机制。其中多头自注意力机制进行了 h 次放缩点积计算, 使模型能够在不同位置联合关注来自不同表示子空间的信息。具体处理过程如下:

$$\text{MultiHead}(Q, K, V) =$$

$$\text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O, \quad (3)$$

其中 $\text{head}_i = \text{Attention}(Q_i, K_i, V_i)$, W^O 是线性投影矩阵。多头自注意力机制进行放缩点积计算时, 会根据头部的个数来动态调整每个头部的维度, 以保持多头自注意力机制的总计算量与全维单头注意力机制的计算量相当, 不增加其计算量。

为减少原始自注意力模块的计算复杂度, 本文采用了分块处理, 具体结构如图 2 所示。在自注意力网络中通过卷积网络映射出不同的空间, 将特征图分成不同的特征块, 将每个特征块的特征空间划分为成不同的特征块, 之后计算当前图像的注意力。分块处理即将 $x \in H \times W \times C$ 的图像进行处理, 变成一系列 $x_p \in N \times (P^2 \cdot C)$ 的展平的 2D 块的序列^[27]。这个序列共有 $N = HW/P^2$ 个展平的 2D 块, 每个块的维度是 $(P^2 \cdot C)$,

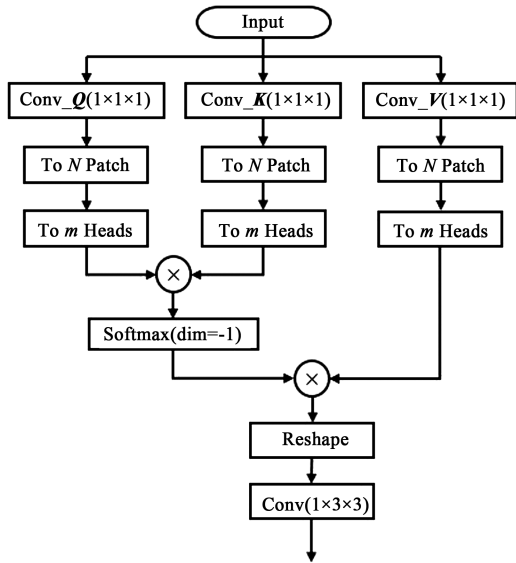


图 2 轻量型多头自注意力机制

Fig. 2 Lightweight multi-head self-attention (L-MHSA)

其中 P 是块大小, C 是通道数。进行分块处理, 可以使网络更关注低层次特征, 提高对低层次特征的提取能力, 同时局部图像变小, 可以降低计算量, 提高模型速度。

2.2.2 残差前馈网络及残差感知单元

残差网络是由一系列残差块组成的。残差块的引入可以解决网络退化现象, 在增加网络层数时, 网络不会发生退化, 提高了网络的性能。残差块的一个通用表示为:

$$x_{l+1} = x_l + F(x_l, W_l), \quad (4)$$

其分为映射和残差两部分, 恒等映射通过直连实现, 残差部分通常由 2~3 个卷积组成。其中直连是深度残差网络中旁路的支线, 可以跳过一层或多层的连接, 直接将输入连接到后面的网络层, 使得后面的网络层可以直接学习残差, 保留了信息的完整性和有效性。这种连接方式不会增加额外的参数计算复杂度, 能够简化学习目标和难度, 加速深度神经网络的训练, 提高深度网络的性能。

本文提出的残差前馈网络在原始的前馈网络基础上添加残差连接, 并使用 LeakyRelu 作为激活函数, 通过改变连接方式来获得更好的性能。具体结构如图 3(a) 所示, 包含了一个扩张层、深度卷积、投影层以及残差连接结构。

视觉任务中经常使用旋转与平移来增广数据, 这些操作应当不能影响模型最终的结果。然而, Transformer 中的绝对位置编码会破坏该不变性。此外, Transformer 忽略了块内的局部相关性与结构信息。为缓解该限制, 本文提出了残差感知单元以提升局部信息, 定义如下:

$$\text{RPU}(X) = \text{DWConv}(X) + X. \quad (5)$$

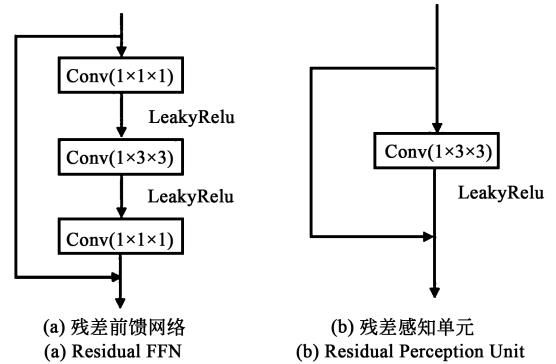


图 3 残差前馈网络与残差感知单元

Fig. 3 Residual FFN and residual perception unit

具体结构如图 3(b) 所示, 由一个卷积核为 $1 \times 3 \times 3$ 的 3D 卷积和一个残差连接组成。

2.3 解码模块

解码模块由反卷积的叠加并添加残差连接构成, 具体结构如图 4 所示。先将残差卷积注意力模块的输出进行转换处理, 从 T 帧变成单帧进行处理, 然后进行反卷积, 卷积核大小为 3×3 , 步长为 1, 将通道数减半, 输入帧特征比例加倍, 之后进行 3×3 卷积, 步长为 1, 将通道数减半, 最后进行 1×1 卷积进行投影。

RCAN 共包含 5 个解码模块, 每个模块通过反卷积将特征反向解码, 将输入帧特征比例依次变为 $1/32, 1/16, 1/8, 1/4$, 最后经过 Tanh 激活函数得到完整的修复图像。其中第 2, 3, 4 个解码模块与对应的残差卷积注意力模块添加了残差连接, 将上一个解码模块的输出与残差卷积注意力模块的输出拼接在一起, 然后输入下一个解码模块。此外第一个解码模块与第五个解码模块的通道数有所不同, 第一个解码模块通道数减半一次, 由 512 降为 256, 而第五个解码模块通道数由 32 降为 16 再降为 3。

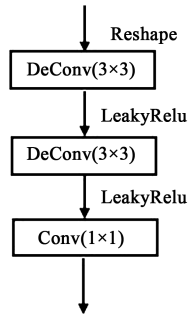


图 4 解码模块

Fig.4 Decode block

2.4 损失函数

选择优化目标函数的原则是确保生成的视频的像素重建精度、修复内容的合理性及时空一致性。该文选择像素级重建损失和时空对抗损失作为损失函数。在生成帧和原始帧之间计算 L_1 , 以此衡量每个像素的重建精度。目标修复区域的 L_1 表示为:

$$L_{rh} = \frac{\|M_1^T \odot (Y_1^T - \hat{Y}_1^T)\|_1}{\|M_1^T\|_1}, \quad (6)$$

原始无空洞的有效区域的 L_1 表示为:

$$L_{rv} = \frac{\|(1 - M_1^T) \odot (Y_1^T - \hat{Y}_1^T)\|_1}{\|1 - M_1^T\|_1}, \quad (7)$$

其中 $\odot$ 表示按元素相乘, 值按照相应区域大小进行归一化, $M_1^T = \{M_1, M_2, \dots, M_T\}$ 表示相应的逐帧掩膜, $Y_1^T = \{Y_1, Y_2, \dots, Y_T\}$ 是原始视频帧, $\hat{Y}_1^T = \{\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_T\}$ 是模型生成的视频帧。

由于对抗性训练有助于高质量内容生成, 该文借鉴使用了 T-PatchGAN (Temporal PatchGAN)^[28] 作为判别器。具体来说, T-PatchGAN 由 6 层 3D 卷积层组成, 其可以学习区分每个时空特征的真假, 以便网络更好地利用真实数据的时空信息和局部到全局的感知细节进行重建。T-PatchGAN 判别器的详细优化函数如下:

$$L_D = E_{x \sim P_{Y_1^T}(x)} [\text{ReLU}(1 - D(x))] + E_{z \sim P_{\hat{Y}_1^T}(z)} [\text{ReLU}(1 + D(z))], \quad (8)$$

本文在其基础上进行改进, 提出了时空对抗损失函数, 具体计算过程如下所示:

$$L_{adv} = -E_{z \sim P_{\hat{Y}_1^T}(z)} [D(z)], \quad (9)$$

结合两部分的总体损失函数为:

$$L = \lambda_{rh} \cdot L_{rh} + \lambda_{rv} \cdot L_{rv} + \lambda_{adv} \cdot L_{adv}, \quad (10)$$

其中: λ_{rh} 、 λ_{rv} 、 λ_{adv} 表示两个像素级重建损失及时空对抗损失在网络总体损失函数中的比重, 通过调整比重参数, 可以更有针对性地调整网络参数, 提高网络训练的灵活性和网络的泛化能力。一般情况下为了平衡时空对抗损失和重建损失, 将损失权重设置为 $\lambda_{rh} = 0.5$, $\lambda_{rv} = 0.5$, $\lambda_{adv} = 0.9$ 。

2.5 训练细节

该网络使用 YouTube-VOS^[29] 数据集进行训练, 训练使用数据集原始的数据分割方式, 其中训练视频 3 471 个, 验证视频 474 个, 测试视频 508 个。将视频帧大小调整为 384×192 , 然后对每个视频帧进行固定掩膜处理, 之后对视频帧进行随机旋转, 提高网络对视频序列旋转角度的鲁棒性, 最后将经过处理后的视频帧输入到网络中进行训练。

该网络使用 4 张 NVIDIA RTX2080Ti 显卡进行训练, 使用 Adam 优化器, 初始学习率为 $1e-4$, batch size 为 6, 步数为 50 万步。此外还可以在 DAVIS^[30] 数据集上训练, 不过由于 DAVIS 数据集规模较小, 一般是将在 YouTube-VOS 数据集上训练得到的模型来初始化模型, 提高其训练速度和精度。

3 实验结果

3.1 实验数据

为了更公平合理地评估本文提出的模型,更好地与现有视频修复模型进行比较,本文采用了视频修复中常用的 DAVIS 和 YouTube-VOS 数据集。其中 DAVIS 是面向实例级分割的数据集,共有 50 个视频,3 455 帧,每个视频序列包含一个对象或者两个空间连接的对象。视频中的每一帧,都拥有像素级别的精度。YouTube-VOS 是迄今为止最大、最全面的视频对象分割数据集。YouTube-VOS 包含 94 种类别,4 453 个 YouTube 视频,每个视频时长约 3~6 s,平均视频长度约为 150 帧,共有 197 272 个对象注释,每个对象由专业注释员手动分割。

本文使用两种自由形式的掩膜进行模型验证:一种是固定掩膜,用来模拟水印去除、空洞修复等应用;另一种是动态掩膜,用来模拟对象移除等应用。固定掩膜是随机生成不规则形状的掩膜并将其应用在整个视频帧的固定区域,动态掩膜是同一对象掩膜应用在整个视频帧中;但每一帧的掩膜区域不同,一般使用数据集中标注的对象分割标注数据。本文使用 YouTube-VOS 数据集进行固定掩膜测试,使用 DAVIS 数据集进行固定掩膜和移动掩膜测试。

3.2 评价标准

实验结果分析采用峰值信噪比 (PSNR)、结构相似性 (SSIM) 和 FID 三种评价标准。其中 PSNR 是基于对应像素点间误差的图像质量评价方法,通过均方差进行定义。其具体计算公式如式 (11) 所示:

$$\text{PSNR} = 10 \cdot \lg \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right), \quad (11)$$

其中 I 是原始图像, K 是生成图像, MAX_I^2 是 I 的最大像素值, MSE 表示 I 与 K 之间的均方误差,计算公式如式 (12) 所示:

$$\text{MSE} = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i, j) - K(i, j)]^2. \quad (12)$$

SSIM 基于原始图像 x 和生成图像 y 之间的亮度、对比度和结构来进行衡量。计算公式如式 (13) 所示:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \quad (13)$$

其中 μ 是均值, σ 是方差, c_1, c_2 为两个常数,避免分母为零。

FID 是计算原始图像和生成图像的特征向量之间距离的一种度量。计算公式如式 (14) 所示:

$$\text{FID}(x, g) = \|\mu_x - \mu_g\| + \text{tr} \left(\sum_x + \sum_g - 2 \left(\sum_x \sum_g \right)^{\frac{1}{2}} \right), \quad (14)$$

其中: x 为原始图像, g 为生成图像, tr 表示矩阵对角线的元素之和, μ 是均值, $\sum$ 是协方差。FID 的数值越小,表示生成图像的多样性和质量越好。

3.3 与现有方法对比

为了更好地验证 RCAN 模型视频修复质量的有效性,本文选择将 RCAN 模型与 4 种现存主流模型进行了对比。其中 VINet^[6]、DFVI^[17]、LGTSM^[31] 是基于卷积神经网络或递归网络的模型,能够同时对所有帧进行修复,但只使用了相邻帧没有使用远距离帧建模。CPN^[22]、OPN^[23] 是基于注意力的模型,对视频逐帧修复,使用了远距离帧建模,而对相邻帧关注度不够。本文提出的 RCAN 模型是在注意力的基础上,使用视频相邻帧及远距离帧即视频所有帧进行建模,能够同时对所有帧进行修复。

本文使用 YouTube-VOS 测试集的 508 个视频,并使用固定掩膜进行模型验证。该网络在 YouTube-VOS 数据集上的视频修复结果与其他模型结果的比较如表 1 所示。与其他模型相比,该模型的 PSNR 为 30.69 dB, SSIM 为 0.965, FID 为 0.059, 运行时间为每帧 0.85 s, 远快于 DFVI 和 LGTSM, 内存为 6 742 MB, 整体性能优于其他模型。结果表明 RCAN 在像素级和整体感知上都有着更好的视频修复质量和较快的修复速度。

为了更好地验证模型性能,实验还比较了不同模型的视觉效果,结果如图 5 所示。图 5(a) 是输入帧,图 5(b) 是 LGTSM 模型的结果,图 5(c) 是 CPN 模型结果,图 5(d) 是本文 RCAN 模型结果。其中第一行和第三行中框选区域为修复区域,第二行和第四行为放大的修复结果,通过对比放大的修复结果可知,RCAN 的修复质量更好,

表 1 RCAN 与现有网络在 YouTube-VOS 上的修复结果对比

Tab.1 Comparison of video inpainting results between RCAN and existing networks on YouTube-VOS

Approachs	PSNR/dB	SSIM	FID	Runtime/s	Memory/MB
VINet ^[6]	28.57	0.914	0.070	0.28	4 110
DFVI ^[17]	28.42	0.913	0.078	112.32	9 794
CPN ^[22]	29.39	0.931	0.068	1.40	9 652
OPN ^[23]	30.28	0.946	0.075	9.05	8 230
LGTSM ^[31]	29.86	0.950	0.076	23.56	8 744
Ours	30.69	0.965	0.059	0.85	6 724

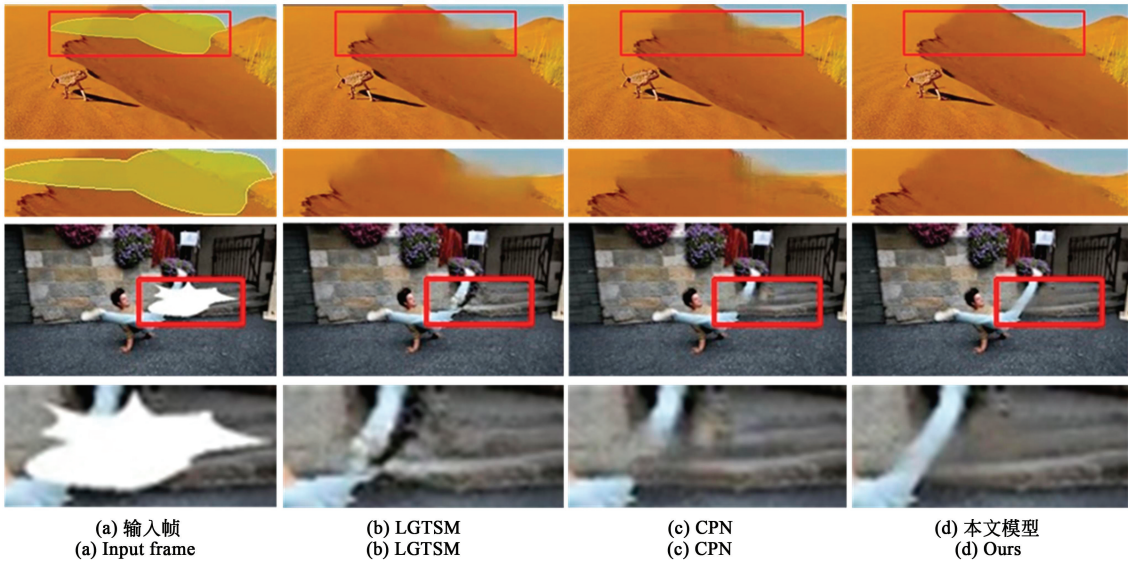


图 5 YouTube-VOS 固定掩膜修复结果

Fig. 5 Video inpainting results for stationary masks on YouTube-VOS

时空结构更一致,修复内容更完整。实验表明了 RCAN 模型在固定掩膜视频修复中的有效性。

此外为了更好地验证网络的通用性,本文还在 DAVIS 数据集上使用固定掩膜以及移动掩膜两种方式进行验证。固定掩膜验证结果如表 2 所

示,PSNR 为 30.67 dB,SSIM 为 0.956,FID 为 0.167,运行时间为每帧 0.69 s,内存消耗为 6 928 MB,整体性能优于其他模型。数据显示在 DAVIS 数据集上,本文模型取得了较高的修复精度和较快的修复速度。

表 2 RCAN 与现有网络在 DAVIS 上的修复结果对比

Tab.2 Comparison of video inpainting results between RCAN and existing networks on DAVIS

Approachs	PSNR/dB	SSIM	FID	Runtime/s	Memory/MB
VINet ^[6]	27.69	0.921	0.193	0.98	4 110
DFVI ^[17]	27.61	0.920	0.186	37.63	9 436
CPN ^[22]	28.16	0.932	0.180	0.83	9 824
OPN ^[23]	30.62	0.946	0.172	4.17	7 868
LGTSM ^[31]	28.15	0.941	0.169	17.45	8 824
Ours	30.67	0.956	0.167	0.69	6 928

使用固定掩膜及移动掩膜两种方式的定性修复结果如图 6 所示,前两行是固定掩膜测试,后两行是移动掩膜测试,图 6(a)是输入帧,图 6(b)是 LGTSM 模型的结果,图 6(c)是

CPN 模型结果。图 6(d)是本文 RCAN 模型结果。对比发现,本文模型结果空间结构相似性和时间连贯性更强,在复杂场景中也有很好的修复结果。



图 6 DAVIS 固定掩膜和移动掩膜修复结果

Fig. 6 Video inpainting results for stationary and moving masks on DAVIS

3.4 消融实验

本文提出的 RCAN 模型将自注意力机制和全局注意力机制引入残差网络中,增强网络对所有帧的空间及时间维度信息的学习能力,保持与相邻帧以及远距离关键帧的时空一致,提高视频修复效果。表 3 是残差网络加入注意力机制后在 DAVIS 数据集上的修复效果对比,修复图像的 PSNR 和 SSIM 得到了大幅度的提升,分别提高了 5.4 dB 和 0.04, FID 降低了 0.075。结果表明引入注意力机制,网络能够关注到重要的特征并抑制无关的干扰特征,能够关注到视频所有帧的关键帧及相邻帧和远距离帧的关键特征,从而提高了模型对所有帧的时空特征学习能力,提高了视频修复效果。

表 3 添加注意力机制修复结果对比

Tab.3 Comparison of adding attention video inpainting results

Approachs	PSNR/dB	SSIM	FID
ResNet	25.27	0.916	0.242
ResNet+注意力机制	30.67	0.956	0.167

此外 RCAN 模型还可以高度自由地定义层数和参数量,通过控制残差卷积注意力模块的残差层数来控制模型深度,以此来控制模型的复杂度和计算量。表 4 是网络在不同残差层数结构的结果比较,数据显示网络层数越多,网络的整体精度越高,但同时也提高了网络的复杂度和计算量。在现实应用中,可根据实际应用情况的不同来定义和选择网络的层数和参数量,提高模型的实际应用能力。

表 4 不同残差层数的 RCAN 在 DAVIS 上的修复结果对比

Tab.4 Comparison of video inpainting results of RCAN with different residual layers on DAVIS

Layers	PSNR/dB	SSIM	FID	Runtime/s	Memory/MB
[1, 2, 2, 1]	29.26	0.937	0.185	0.18	2110
[1, 2, 4, 1]	29.41	0.940	0.181	0.24	2404
[2, 4, 6, 2]	30.38	0.953	0.176	0.46	4394
[3, 4, 6, 3]	30.46	0.954	0.171	0.55	4980
[4, 8, 8, 4]	30.67	0.956	0.167	0.69	6928

4 结 论

本文提出了一种新的用于视频修复的残差卷积注意力网络,通过将自注意力机制和全局注意力机制引入进残差网络中,规避了卷积结构由于其感受野而无法获得全局信息的缺陷,并由此获得更加强大的表达能力,增强了网络对视频输入帧时空特征的学习能力以及泛化能力,提高了视频修复质

量。同时结合时空对抗损失,提升修复内容的真实性,提高了网络的整体性能。此外网络还能够根据现实应用中的实际情况自由地定义层数和参数量,来权衡网络精度和速度,提高了模型的实际应用能力。实验表明,本文提出的模型在 YouTube-VOS 和 DAVIS 数据集上的修复效果明显优于其他模型,对比 DFVI 模型,平均 PSNR 高出 2.67 dB, SSIM 高出 0.044, FID 降低了 12.98%,表明了残差卷积注意力网络在视频修复中的有效性。

参 考 文 献:

- [1] 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. CVPR 2001*. Kauai: IEEE, 2001: 355-362.
- [2] GRANADOS M, TOMPKIN J, KIM K, *et al.* How not to be seen-object removal from videos of crowded scenes [J]. *Computer Graphics Forum*, 2012, 31(2pt1): 219-228.
- [3] 徐展,曹哲.复杂运动摄像机拍摄视频的背景修复技术[J].*计算机应用*, 2014, 34(12): 3540-3544, 3559.
XU Z, CAO Z. Video background completion with complexly moving camera [J]. *Journal of Computer Applications*, 2014, 34(12): 3540-3544, 3559. (in Chinese)
- [4] MATSUSHITA Y, OFEK E, GE W N, *et al.* Full-frame video stabilization with motion inpainting [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006, 28(7): 1150-1163.
- [5] PATWARDHAN K A, SAPIRO G, BERTALMIO M. Video inpainting of occluding and occluded objects [C]//*Proceedings of the IEEE International Conference on Image Processing 2005*. Genova: IEEE, 2005: II-69.
- [6] KIM D, WOO S, LEE J Y, *et al.* Deep video inpainting [C]//*Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach: IEEE, 2019: 5785-5794.
- [7] TANG N C, HSU C T, SU C W, *et al.* Video inpainting on digitized vintage films via maintaining spatiotemporal continuity [J]. *IEEE Transactions on Multimedia*, 2011, 13(4): 602-614.
- [8] KIM D, WOO S, LEE J Y, *et al.* Deep blind video decaptioning by temporal aggregation and recurrence [C]//*Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach: IEEE, 2019: 4258-4267.
- [9] HUANG J B, KANG S B, AHUJA N, *et al.* Temporally coherent completion of dynamic video [J]. *ACM Transactions on Graphics*, 2016, 35(6): 196.
- [10] 付傲威,赵敏,罗令,等.自由立体显示中基于深度卷积神经网络的虚拟视点生成方法[J].*液晶与显示*, 2019, 34(11): 1031-1036.

- FU A W, ZHAO M, LUO L, *et al.* Virtual viewpoint image generation method based on deep convolutional neural network in autostereoscopic display [J]. *Chinese Journal of Liquid Crystals and Displays*, 2019, 34(11): 1031-1036. (in Chinese)
- [11] NEWSON A, ALMANSA A, FRADET M, *et al.* Video inpainting of complex scenes [J]. *SIAM Journal on Imaging Sciences*, 2014, 7(4): 1993-2019.
- [12] SHIH T K, TANG N C, HWANG J N. Exemplar-based video inpainting without ghost shadow artifacts by maintaining temporal continuity [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2009, 19(3): 347-360.
- [13] 吴磊,吕国强,赵晨,等.基于多尺度残差网络的CT图像超分辨率重建[J].液晶与显示,2019,34(10):1006-1012.
WU L, LYU G Q, ZHAO C, *et al.* CT image super-resolution reconstruction based on multi-scale residual network [J]. *Chinese Journal of Liquid Crystals and Displays*, 2019, 34(10): 1006-1012. (in Chinese)
- [14] 陈宗航,胡海龙,姚剑敏,等.基于改进生成对抗网络的单帧图像超分辨率重建[J].液晶与显示,2021,36(5): 705-712.
CHEN Z H, HU H L, YAO J M, *et al.* Single frame image super-resolution reconstruction based on improved generative adversarial network [J]. *Chinese Journal of Liquid Crystals and Displays*, 2021, 36(5): 705-712. (in Chinese)
- [15] 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 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul: IEEE, 2019: 9065-9074.
- [16] WANG C, HUANG H B, HAN X G, *et al.* Video inpainting by jointly learning temporal structure and spatial details [J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, 33(1): 5232-5239.
- [17] XU R, LI X X, ZHOU B L, *et al.* Deep flow-guided video inpainting [C]//*Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach: IEEE, 2019: 3718-3727.
- [18] ZHANG H T, MAI L, JIN H L, *et al.* An internal learning approach to video inpainting [C]//*Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul: IEEE, 2019: 2720-2729.
- [19] YU J H, LIN Z, YANG J M, *et al.* Free-form image inpainting with gated convolution [C]//*Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul: IEEE, 2019: 4470-4479.
- [20] LIN J, GAN C, HAN S. TSM: temporal shift module for efficient video understanding [C]//*Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul: IEEE, 2019: 7082-7092.
- [21] 于冰,丁友东,谢志峰,等.基于时空生成对抗网络的视频修复[J].计算机辅助设计与图形学学报,2020,32(5): 769-779.
YU B, DING Y D, XIE Z F, *et al.* Temporal-spatial generative adversarial networks for video inpainting [J]. *Journal of Computer-Aided Design & Computer Graphics*, 2020, 32(5): 769-779. (in Chinese)
- [22] LEE S, OH S W, WON D, *et al.* Copy-and-paste networks for deep video inpainting [C]//*Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul: IEEE, 2019: 4412-4420.
- [23] OH S W, LEE S, LEE J Y, *et al.* Onion-peel networks for deep video completion [C]//*Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul: IEEE, 2019: 4402-4411.
- [24] HE K M, ZHANG X Y, REN S Q, *et al.* Deep residual learning for image recognition [C]//*Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas: IEEE, 2016: 770-778.
- [25] VASWANI A, SHAZEER N, PARMAR N, *et al.* Attention is all you need [C]//*Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach: Curran Associates Inc., 2017: 6000-6010.
- [26] HAUSMAN D M, WOODWARD J. Independence, invariance and the causal markov condition [J]. *The British Journal for the Philosophy of Science*, 1999, 50(4): 521-583.
- [27] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al.* An image is worth 16×16 words: transformers for image recognition at scale [J]. *arXiv preprint arXiv:2010.11929*, 2020.
- [28] YANG F Z, YANG H, FU J L, *et al.* Learning texture transformer network for image super-resolution [C]//

- Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle: IEEE, 2020: 5790-5799.
- [29] CAELLES S, MONTES A, MANINIS K K, *et al*. The 2018 DAVIS challenge on video object segmentation [J]. *arXiv preprint arXiv:1803.00557*, 2018.
- [30] XU N, YANG L J, FAN Y C, *et al*. YouTube-VOS: a large-scale video object segmentation benchmark [J]. *arXiv preprint arXiv:1809.03327*, 2018.
- [31] CHANG Y L, LIU Z Y, LEE K Y, *et al*. Learnable gated temporal shift module for free-form video inpainting [C]//*Proceedings of the 30th British Machine Vision Conference*. Cardiff, UK: BMVA Press, 2019: 149.

作者简介:



李德财(1997—),男,山东枣庄人,硕士研究生,2019 年于聊城大学获得学士学位,主要从事视频语义理解方面的研究。E-mail: n191127093@fzu.edu.cn



严 群(1965—),男,上海人,博士,教授,1995 年于美国范德堡大学获得物理学博士学位,主要从事 Micro-LED 显示技术、场景视频和交互行为采集方面的研究。E-mail: qunfyan@gmail.com