

# 深度视频修复篡改的被动取证研究

熊义毛<sup>1</sup>, 丁湘陵<sup>1,3,5</sup>, 谷庆<sup>1</sup>, 杨高波<sup>2</sup>, 赵险峰<sup>3,4</sup>

<sup>1</sup> 湖南科技大学 计算机科学与工程学院 湘潭 中国 411201

<sup>2</sup> 湖南大学 信息科学与工程学院 长沙 中国 410082

<sup>3</sup> 中国科学院信息工程研究所 信息安全国家重点实验室 北京 中国 100093

<sup>4</sup> 中国科学院大学 网络空间安全学院 北京 中国 100093

<sup>5</sup> 郑州信大先进技术研究院 郑州 中国 450000

**摘要** 深度视频修复技术就是利用深度学习技术,对视频中的缺失区域进行补全或移除特定目标对象。它也可用于合成篡改视频,其篡改后的视频很难通过肉眼辨别真假,尤其是一些恶意修复的视频在社交媒体上传播时,容易造成负面的社会舆论。目前,针对深度视频修复篡改的被动检测技术起步较晚,尽管它已经得到一些关注,但在研究的深度和广度上还远远不够。因此,本文提出一种基于级联 ConvGRU 和八方向局部注意力的被动取证技术,从时空域角度实现对深度修复篡改区域的定位检测。首先,为了提取修复区域的更多特征,RGB 帧和错误级分析帧 ELA 平行输入编码器中,通过通道特征级融合,生成不同尺度的多模态特征。其次,在解码器部分,使用编码器生成的多尺度特征与串联的 ConvGRU 进行通道级融合来捕捉视频帧间的时域不连续性。最后,在编码器的最后一级 RGB 特征后,引入八方向局部注意力模块,该模块通过八个方向来关注像素的邻域信息,捕捉修复区域像素间的异常。实验中,本文使用了 VI、OP、DSTT 和 FGVC 四种最新的深度视频修复方法与已有的深度视频修复篡改检测方法 HPF 和 VIDNet 进行了对比,性能优于 HPF 且在编码器参数仅 VIDNet 的五分之一的情况下获得与 VIDNet 可比的性能。结果表明,本文所提方法利用多尺度双模态特征和引入的八方向局部注意力模块来关注像素间的相关性,使用 ConvGRU 捕捉时域异常,实现像素级的篡改区域定位,获得精准的定位效果。

**关键词** 深度视频修复; 视频篡改检测; 级联 ConvGRU; 局部注意力模块; 空时预测

**中图分类号** TP309 **DOI号** 10.19363/J.cnki.cn10-1380/tn.2024.07.08

## The Passive Forensics of Deep Video Inpainting

XIONG Yimao<sup>1</sup>, DING Xiangling<sup>1,3,5</sup>, GU Qing<sup>1</sup>, YANG Gaobo<sup>2</sup>, ZHAO Xianfeng<sup>3,4</sup>

<sup>1</sup> School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan 411201, China

<sup>2</sup> College of Computer and Communication, Hunan University, Changsha 410082, China

<sup>3</sup> State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China

<sup>4</sup> School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100093, China

<sup>5</sup> Zhengzhou Xinda Institute of Advanced Technology, Zhengzhou 450000, China

**Abstract** Deep video inpainting is to fill missing areas or remove the specific target objects in the video by using deep learning technology. It is also exploited to synthesize tampered videos. The tampered videos are arduous to be identified with the naked eye. Especially, it is easy to cause negative public perspectives when some maliciously inpainted videos are spread on social media. At present, although it has received some attentions, its passive detection was far from enough in the depth and breadth of research for deep video inpainting. Therefore, this paper proposes a passive forensics technique based on cascaded ConvGRU and eight-direction local attention to achieve the localization of inpainted regions in deep tampered videos. The proposed method aims to localize the tampered regions in deep inpainted videos from the spatio-temporal domain. Firstly, RGB frames and error-level analysis frames, ELA, are fed into the encoder in parallel to extract more features of the inpainted area, and then multi-modal features are generated at different scales through channel feature-level fusion. Secondly, in the decoder, encoder-generated multimodal features cascaded ConvGRUs are utilized to capture the temporal continuity between video frames. Finally, in the last level RGB feature of the encoder, an eight-direction local attention module is introduced, which pays attention to the neighborhood information of pixels through eight directions and captures the anomaly between pixels in the inpainted area. In the experiment, four latest deep video inpainting methods, VI, OP, DSTT, and FGVC, were used to compare their performance with existing deep video inpainting tamper detection methods, HPF and VIDNet. The performance was superior to HPF and comparable to VIDNet was achieved when the encoder parameters were only one-fifth of VIDNet. The results show that the proposed method

**通讯作者:** 丁湘陵, 博士, 教授, Email: xianglingding@hnust.edu.cn.

本课题得到国家自然科学基金(No. 62272160); 信息安全国家重点实验室开放课题(No. 2021-ZD-07); 河南省网络空间态势感知重点实验室开放课题基金资助(No. HNTS2022025)资助。

收稿日期: 2022-10-31; 修改日期: 2023-04-12; 定稿日期: 2024-04-07

focuses on the correlation between pixels by generating multi-modal features and introduces an eight-direction local attention module. Concurrently, the ConvGRU takes advantage of capturing temporal anomalies, by achieving tampered positioning, and obtaining accurate localization effect.

**Key words** deep video inpainting; video forgery detection; cascaded ConvGRU; local attention module; spatio-temporal prediction

## 1 引言

视频修复技术, 就是利用视频帧的周围信息或是相似帧对视频中的缺失区域进行补全。近年来, 已经引起了计算机视觉领域的关注。随着多媒体技术的快速发展, 视频修复技术变得越发成熟, 尤其是最近的深度视频修复方法<sup>[1-4]</sup>, 其修复后的视频看起来自然流畅, 让人无法用肉眼看出修复痕迹。然而, 视频修复技术在带给人们便捷生活的同时, 也有可能带来一些不良影响。因为这些视频修复技术很可能会被恶意使用。例如, 一些不法分子在社交平台上发布一些虚假信息, 从而获得点击量、引发负面的社会舆论。更有甚者, 部分犯罪分子将修复后的视频作为司法取证的证据, 以此来动摇陪审团, 使其逃脱法律的制裁。因此, 针对视频修复篡改的被动取证技术的发展已经变得刻不容缓。而本文的目标就是开发一个可用于深度修复篡改视频的检测框架, 效果演示如图 1 所示。其中, 第一列为原始视频, 第二列为采用 Oh S W 等人<sup>[1]</sup>提出的深度修复方法移除骆驼后的修复视频, 第三列为采用本文提出方法在时间和空间上输出的篡改预测值, 第四列为篡改掩膜。从图 1 中, 我们通过肉眼很难发现第二列的修复痕迹, 而本文提出的方法能有效的定位出篡改区域且具有较好的边缘细节。

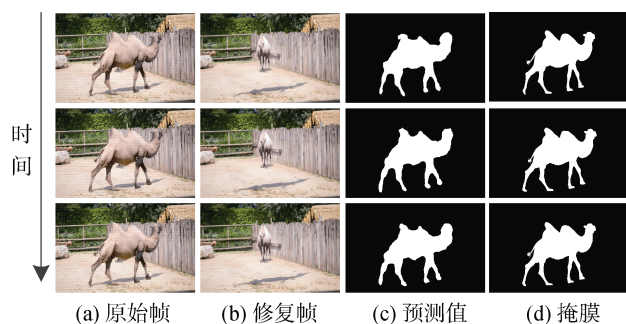


图 1 深度修复篡改视频的检测图例

Figure 1 Detection legend of deep inpainting tampered video

目前为止, 有关视频修复检测的方法已经出现很多, 主要包括传统方法和基于深度学习的方法。对于传统方法而言, 视频修复依靠视频帧的边缘信息

或视频帧之间的相似信息来对视频进行修复, 因此会留下一些修复痕迹, 如伪影和边缘失真等。而在视频伪造检测中, 检测器通常依靠这些视频修复痕迹, 从而能获得较好的检测效果。但当采用深度视频修复方法时, 这些传统视频修复遗留的痕迹会被弱化甚至被抹除, 因此导致这些检测器的性能出现退化。

由于深度学习超强的特征表征能力, 基于深度学习的深度视频修复被动取证方法应运而生。其中, 最有代表性的是 Zhou P 等人<sup>[5]</sup>提出的深度视频修复检测(Deep Video Inpainting Detection, VIDNet)。该方法采用 RGB 帧和错误级分析帧(Error Level Analysis, ELA)作为输入, 使用 VGG16 作为编码器进行特征提取, 采用四方向局部注意力模块捕捉像素点上下左右的空域一致性, 使用 ConvLSTM 对视频帧进行时空序列预测, 实现区域级定位。其方法虽能得到很好的检测效果, 但像素间除了上下左右, 还应该在对角和次对角上也具有强相关<sup>[6]</sup>; 且 ConvGRU 由于更少的参数和高计算效率, 具有更高效的时域信息捕捉能力。随后, Ding 等人<sup>[7]</sup>提出了一种基于八方向注意和级联 ConvGRU 的深度视频修复篡改取证算法, 具有更好的纹理细节捕捉能力。本文在此基础上进一步扩充, 采用路径聚合网络(Path Aggregation Network, PANet)进行编码器优化, 提取特征, 并通过通道级特征融合, 在编码器的最后一级引入八方向局部注意力模块来关注像素的周围相关信息, 最后利用 ConvGRU 捕捉时域信息, 对修复视频进行时空序列预测。在利用 Davis 数据集构建的原始视频和修复视频数据集上, 提出的方法能有效的检测到深度视频修复区域, 且具有较好的边缘细节信息。此外, 本文还对该方法进行了实验扩展, 即对数据集进行压缩、增加高斯噪声等后处理操作, 然后进行训练与测试, 以此来检测模型的性能。最后, 本文提出的方法在编码器以仅 VIDNet 五分之一的参数获得与之相近的性能。

本文贡献主要有四个方面: 1)提出了一种针对深度视频修复篡改的检测框架, 该框架能够很好的提取视频中的时空篡改区域; 2)使用 RGB 帧和错误级分析帧作为平行输入, 利用编码器进行特征提取并在不同尺度上进行特征级联融合; 3)引入八方向局部注意力模块关注像素的周围相关信息; 4)使用 ConvGRU

对视频进行时空预测, 捕捉视频的时域一致性。

## 2 相关工作

### 2.1 深度视频修复篡改技术

近年来, 图像修复技术的进步推动着视频修复技术的发展<sup>[8-11]</sup>。目前, 视频修复技术已经越发成熟与稳重, 主要分为两大类, 基于补丁和基于学习的方法。其中, 基于补丁的方法主要依靠视频帧的周围信息或是相似信息来对视频帧的缺失区域进行补全。Barnes 等人<sup>[12]</sup>提出了一种最为典型的方法, 主要通过卷积迭代式的搜索篡改区域的边缘信息来得到更多相似信息, 以此来对篡改区域的进行修复。为了让修复效果更好, Huang 等人<sup>[13]</sup>提出了一种优化算法, 通过联合估计缺失区域中的光流和颜色, 用时间连贯的方式来合成视频中的缺失区域。而基于学习的方法主要通过网络模型不停地反向回归, 寻找最优参数, 以此来达到视频修复效果。Kim D 等人<sup>[2]</sup>提出了一种基于编解码器架构, 旨在从相邻帧中提取有效信息来修复篡改区域的深度网络体系结构, 能生成语义上更为准确, 时间上更加流畅的修复视频, 但当篡改区域较大时, 修复效果并不理想。Liu R 等人<sup>[3]</sup>提出了一种解耦时空变换网络, 通过时间解耦器来注意不同帧上对象的时间运动, 通过空间解耦器来注意相似的背景纹理, 在视频补全和目标去除方面取得了很好的性能, 但模型的性能还有待提高。Zeng Y 等人<sup>[14]</sup>提出了一种联合时空变换学习的视频修复方法, 通过自我关注的同时, 填充所有输入帧的缺失区域, 并利用时空对抗损失来优化网络, 表现出了很好的视频重建质量, 但复杂运动的短期时间连续性很难捕捉到。Gao C 等人<sup>[4]</sup>提出了一种基于光流边缘指导的视频补全算法。利用光流估计来提取运动边缘信息, 然后用其来指导带有尖锐边缘的分段平滑流补全, 以此达到视频修复效果, 但方法运行速度慢, 性能较差。Oh S W 等人<sup>[1]</sup>提出了一种基于洋葱皮网络的视频补全方法, 该方法通过参考输入帧, 利用上下文信息从缺失的边界开始填补, 取得了令人满意的修复效果。

### 2.2 视频修复篡改的被动取证

传统的视频篡改检测方法可分为基于帧间差异的检测方法和基于帧内差异的检测方法。基于帧间差异检测的方法主要针对由图像修复方法篡改的修复视频, 它的伪造痕迹出发点在于, 由于基于图像修复的视频修复是逐帧进行的, 没有考虑到视频的时空连续性, 因此修复后的视频往往会出现时间不连续, 空间不一致的现象。其代表性的方法有: 李倩

等人<sup>[15]</sup>提出了一种基于视频修复的运动目标删除篡改行为的检测算法。对于篡改后未压缩视频采用对称帧差法检测, 对于篡改后压缩视频通过运动光流场来进行检测。白珊山等人<sup>[16]</sup>提出了一种基于双通道卷积神经网络的视频目标移除取证算法。利用双线性池化来对 RGB 特征和噪声特征进行融合, 从而有效地识别篡改帧。熊潇等人<sup>[17]</sup>提出了一种基于预测残差检测的数字视频篡改鉴定方法。利用预测残差出现的周期性特征, 来检测视频是否经过帧删除或帧插入操作。黄添强等人<sup>[18]</sup>提出了一种利用内容连续性的数字视频篡改检测算法。先计算视频帧间相关性的变换度, 再利用切比雪夫不等式自适应设定阈值, 判断出离群点。Amrini I 等人<sup>[19]</sup>提出了一种基于 CNN 的光流深度伪造视频检测方法, 利用光流场差异来鉴别篡改视频与真实视频。Güera D 等人<sup>[20]</sup>提出了一种 CNN 和长短期记忆网络相结合的方法来检测深度伪造视频。Saxena S 等人<sup>[21]</sup>提出了一种基于光流不一致的视频修复检测与时域定位的方法。先通过光流矩阵来区分真实视频与修复视频, 再利用光流矩阵建模马尔科夫链提取转移概率矩阵 (Transition Probability Matrix, TPM) 来进行基于支持向量机的分类。以此来判断真实帧与篡改帧。基于帧内差异检测的方法主要针对篡改区域与真实区域无法正常拟合的修复视频。其伪造视频往往是通过提取未篡改区域的像素信息来对篡改区域进行的修复, 因此修复区域与真实区域会出现边缘效应和视觉伪影现象。而基于帧内差异的检测方法便能很好的应对此种现象, 检测方法主要有: Zhu Y 等人<sup>[22]</sup>提出了一种基于复制移动篡改检测的自适应注意和残差精化网络 AR-Net。利用自适应注意力机制和通道注意力特征来充分捕捉上下文信息, 通过残差精化网络对粗掩膜进行细化, 生成最终掩膜。Hsu C C 等人<sup>[23]</sup>提出了一种基于噪声残差相关性的视频修复篡改检测网络。通过建模高斯混合模型来定位篡改帧。Lin C S 等人<sup>[24]</sup>提出了一种基于时空相干分析的区域级视频伪造被动检测与定位方法。利用时空切片的方法来检测视频在连续帧中是否遭受异常高或异常低的相干性, 以此来定位篡改区域。Bagiwa M A 等人<sup>[25]</sup>提出了一种基于 Hessian 矩阵相关性的数字视频修复检测方法。利用海森矩阵 (Statistical Correlation of Hessian Matrix, SCHM) 的统计相关性来检测和定位视频序列中的篡改区域。袁秀娟等人<sup>[26]</sup>提出了一种基于边缘异常与压缩跟踪的视频抠像篡改检测方法。通过对视频帧进行 Sobel 边缘检测来确定篡改区域。张静等人<sup>[27]</sup>提出了一种基于滤波检测的视频区域篡改检测算法。通过滤波处理对视频帧高频能

量的影响来构建相应的篡改检测特征。翁韶伟等人<sup>[28]</sup>提出一种基于 Inception-V3 网络的双阶段数字视频篡改检测算法。通过一级分类器来区分原始视频和篡改视频, 利用二级分类器来定位篡改区域。陈临强等人<sup>[29]</sup>提出了一种基于视频对象移除篡改的时空域定位被动取证算法。将特征提取器与鉴别器结合来定位视频篡改时域, 将特征提取器与空域定位器相结合来定位视频篡改空域。Li A 等人<sup>[30]</sup>提出了走向深度修复的普遍检测方法。通过设计噪声图像交叉融合网络来有效的定位篡改区域。

随着深度学习的出现, 视频修复篡改技术变得越发成熟, 修复痕迹微弱。因此, 传统的视频修复篡改检测方法已不能有效检测这类篡改手段。鉴于深度学习技术较强的特征表征能力, 基于深度学习的视频修复检测方法能很好的弥补传统方法所带来的不足。如 Peng Zhou 等人<sup>[5]</sup>提出了一种基于深度视频修复检测网络, 是一种带有注意模块的双流编解码器架构, 具有很好的定位效果, 但其在篡改边缘的细节定位上有待提升; 深圳大学的 H Li 等人<sup>[31]</sup>提出了一种利用高通卷积网络实现深度图像修复定位 (Localization of Deep Inpainting Using Hight-Pass Fully Convolutional Network, HPF), 其定位效果精准,

但在深度视频修复检测上缺乏视频特有的时域信息的捕捉; X Ding 等人<sup>[32]</sup>提出了一种基于时空卷积和细化网络的深度视频修复定位; Yu B 等人<sup>[33]</sup>提出了一种基于频率感知的时空转换器用于视频修复检测。这些方法要么都在单个模态下提取特征, 且没有关注像素间八个方向的相关性, 要么忽视视频特有的时域信息。本文利用 RGB 帧和错误级分析帧 ELA 的互补性, 通过通道级特征融合, 结合八方向局部注意力模块, 有效捕捉空时域信息, 实现深度视频修复篡改区域的像素级定位。

### 3 本文方法

如图 2 所示, 本文设计了一个端到端的深度视频修复定位框架。该框架利用三个基本模块来完成定位任务: 首先, 编码器模块采用两个并行的 PANet 网络从连续的 RGB 帧和相应的 ELA 帧中学习鉴别特征, 并进行通道特征级融合; 然后, 引入八方向局部注意力模块来关注篡改区域的像素间的相关性, 使得提取的特征的语义信息更加丰富; 最后, 级联八个 ConvGRU 和上采样层到解码器模块实现像素级的时空定位。

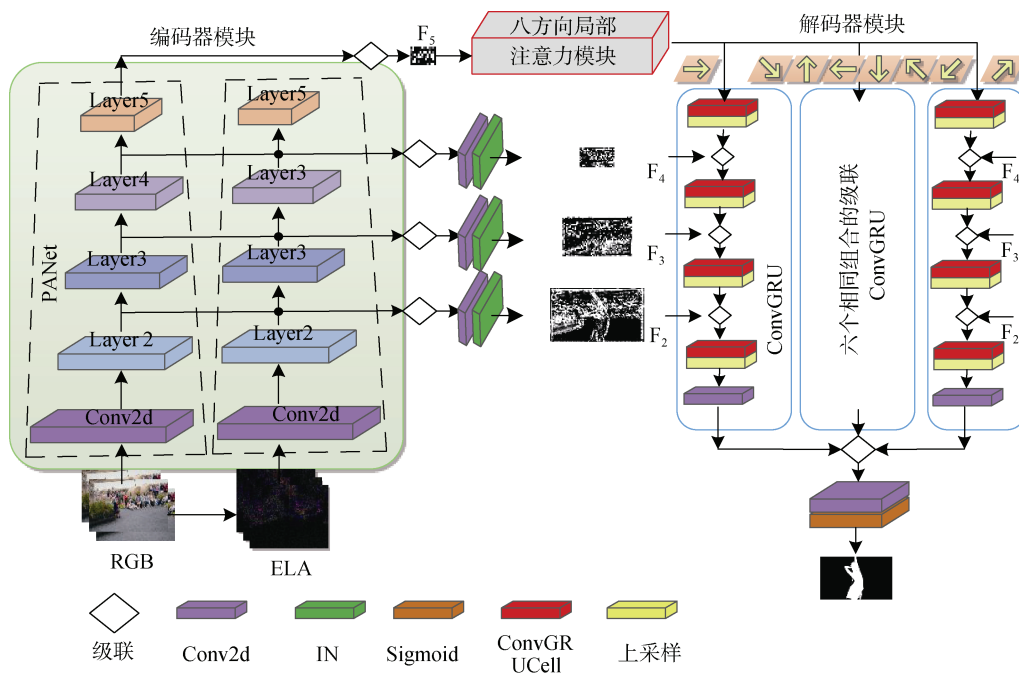


图 2 深度视频修复篡改的被动检测网络模型图

Figure 2 The network framework of pixel-level tampering localization for deep video inpainting

#### 3.1 增强的篡改痕迹

捕捉篡改操作留下的痕迹是实现视频取证的一项基本任务。由于深度视频修复的内容在视觉上难以分辨, 因此有必要对篡改痕迹进行增强。在现有的

取证工作中, 误差水平分析<sup>[5]</sup>和八方向操作<sup>[6]</sup>是增强被修复区域与其周围像素之间关系的常用预处理操作手段。受这些工作的启发, 我们计划利用 ELA 和八方向操作增强深度视频修复篡改所遗留的伪造痕



迹。ELA 的计算如公式(1)所示:

$$f_{ELA} = |f - f_{code}| \quad (1)$$

其中  $|\cdot|$  表示绝对值运算,  $f_{ELA}$  表示 ELA 帧,  $f$  表示原始帧,  $f_{code}$  表示从编码帧中重建的解码帧。

本文采用 TPM 作为视频帧的统计量, 并将篡改视频与真实视频进行了对比, 细分为三种方案, 以此来验证八方向的性能。在第一种方案中, 统计量直接计算 RGB 篡改视频与真实视频之间的差异; 第二种方案中, 统计量计算 ELA 篡改视频与真实视频之间的差异; 第三种方案, 统计量计算篡改视频与真实视频在经过八方向局部注意模块后的差异。假设一个目标矩阵是 RGB 帧, ELA 或提取的特征, 灰度级有  $N$  层, 则相邻位置的 TPM 计算如公式(2)所示:

$$TPM_{x,y} = \Pr(T_{i,j+1} = y | T_{i,j} = x), 1 \leq x, y \leq N \quad (2)$$

其中  $(i, j)$  和  $(x, y)$  分别表示元素在目标数组  $T$  中的位置和 TPM 中的位置。为了进行实验分析, 一些来自 Davis<sup>[34]</sup>的视频使用篡改方法 VI<sup>[2]</sup>对所提供的掩模进行视频修复。然后根据这些原始视频区域和修复的区域计算 TPM。平均 TPM 分别显示在图 3 (a) 和(b)中。对 ELA 帧以及经过八方向局部注意模块后的特征也执行了类似的计算。对应的平均 TPM 分别如图 3 (c)、(d)和(e)、(f)所示。由于保存图片时, 需要将像素值反归一化到  $[0, 255]$ , 所以 RGB 帧和 ELA 帧的像素值范围为  $[0, 255]$ 。而在八方向局部注意模块中, 特征值经过 sigmoid 处理后变为  $[0, 1]$  之间的数值, 在对其进行反归一化保存图片, 其像素值范围也是  $[0, 255]$ 。因此, 上图中所有图都具有相同的灰色强度。

从图 3 可以观察到, 来自 RGB 帧的 TPM 在修复区域与原始区域之间表现出更多的相似性, 而来自 ELA 帧的 TPM 与经过八方向局部注意模块后的特征分别在矩形、椭圆上表现出明显的差异。究其原因, 可能是修复操作主要利用周围的像素来填充视觉逼真内容的移除区域, 而未能保持自然视频中固有存在的关系。由此, 可以得出结论, ELA 框架和八方向局部注意模块可以增强修复痕迹。

### 3.2 PANet 模块

一般来说, 网络的底层特征中包含更多的位置信息, 高层特征中包含着更多的语义信息, 而特征金字塔(Feature Pyramid Networks, FPN)<sup>[35]</sup>采用了自上而下、自下而上的方式, 将深层信息进行上采样, 与浅层信息逐元素相加, 并通过横向连接来将上采样后的高层语义特征与浅层的定位细节特征进行融合, 从而构建了尺度不同的金字塔结构, 以此来传

播语义上较强的特征, 如图 4(a)所示。然而, 这种自顶向下的模式, 虽然能将高层特征传下来, 但底层特征却无法影响高层特征, 对定位信息没有传递, 且是逐层传递, 计算量比较大, 有助于检测对象较大的物体, 对小对象检测来说并不友好。因此, PANet 在此基础上增加了自底向上的路径(如图 4(b))

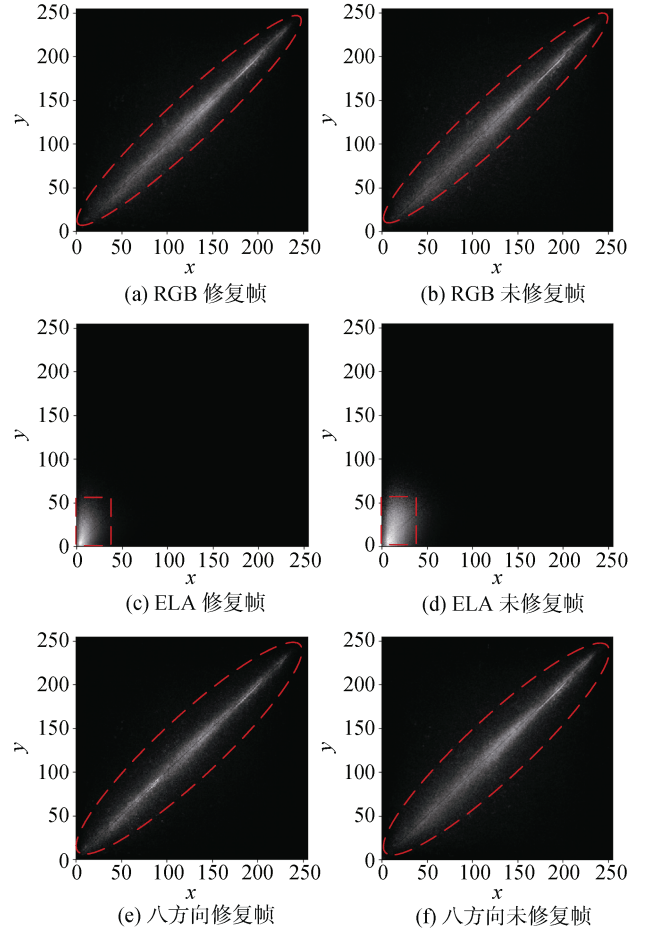


图 3 修复视频与未修复视频的概率转移矩阵图  
Figure 3 Probability transition matrix diagram of inpainted videos and original videos

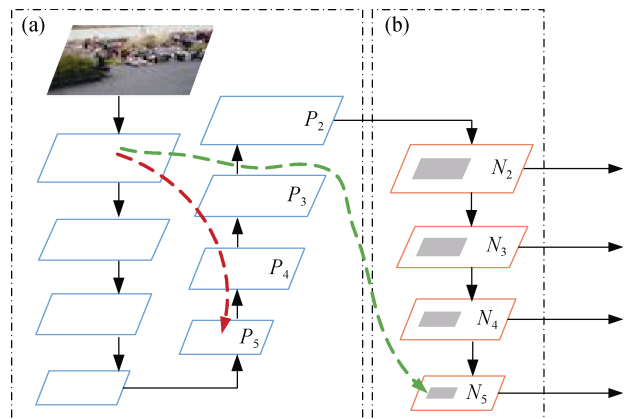


图 4 PANet 网络  
Figure 4 Network of PANet

所示), 将底层的强定位特征传递上去, 提高底层信息的利用率, 缩短较低层与最上层特征之间的信息路径, 使得底层信息更容易传递到高层顶部。在图 4 中, 红线是 FPN 中浅层信息的传递路径, 至少经过上百个网络层, 因此很可能会出现细节信息丢失的问题, 增加检测难度。而绿线是 PANet 的浅层信息传递路径, 经过的路径小于十层, 并且能将所有特征级别的信息积聚起来, 从而缩短了较低层与最高层特征之间的距离, 以实现可靠的信息传递。

### 3.3 具有并行特征提取的编码器模块

虽然 ELA 增强了修复痕迹, 但它们往往会触发块<sup>[5]</sup>边界的误报警, 而块是视频编码器的基本单元。因此, 本文利用了 RGB 帧的互补性。设计的编码模块由两个并行的特征提取器组成, 其中一个使用连续的 RGB 帧作为输入, 另一个接收额外的增强 ELA。每个提取器使用 PANet 来提取区分特征。在每个 PANet 中, 有 5 个连续的卷积层, 这里取后四层卷积层, 将输入转换为不同尺度下的 4 个特征表示。除了第一层的  $7 \times 7$  外, 大多数层的内核大小都是  $3 \times 3$ 。将它们的通道数分别设置为 64、256、512、1024 和 2048, 除了第一层步长为 1, 且去掉下采样操作, 其余步长都设置为 2。在四个多模态特征中, 高层次的特征能获得更丰富的语义信息, 适合小目标检测; 低层次的特征能获得更丰富的轮廓, 适合大目标检测。PANet 具有自底向上的路径, 能将底层的强定位特征传递上去, 提高底层信息的利用率。随后, 对于具有相同尺度的特征图, 首先对它们进行归一化操作, 然后进行通道特征级融合, 再连接一个  $3 \times 3$  的卷积层, 以形成由 RGB 和 ELA 帧组成的统一表示, 如公式(3)所示:

$$\begin{aligned} F_m &= N(W_0 * [F_m^{RGB}, F_m^{ELA}]) \quad m=1, 2, 3, 4 \\ F_m &= [F_m^{RGB}, F_m^{ELA}] \quad m=5 \end{aligned} \quad (3)$$

其中,  $F_m^{RGB}$  和  $F_m^{ELA}$  是尺度为  $m$  的 RGB 和 ELA 特征映射,  $W_0$  是  $3 \times 3$  卷积的权重,  $[\cdot]$ 、 $*$  和  $N(\cdot)$  分别表示特征级联、卷积和归一化操作。

### 3.4 八方向局部注意模块

因为视频修复主要依靠当前帧或相邻帧的周围像素填充来执行的, 所以修复区域中像素之间固有的相关性在一定程度上会不可避免地受到破坏。这是深度修复定位所需要的特征。因此, 受八方向差分技术<sup>[6]</sup>的启发, 将其扩展到八方向局部注意力模块中, 以揭示相邻像素间的相关性。

引入八个方向的局部注意映射, 研究在哪里进

行聚焦能表示对应方向上像素间的相关性, 如图 5 所示。八个方向的注意特征映射分别由上标  $\{\uparrow, \downarrow, \leftarrow, \rightarrow, \nearrow, \searrow, \swarrow, \nwarrow\}$  表示。例如, 左上至右下的注意函数  $Att_5^{\searrow}$ , 以及其他七个注意函数, 即  $(Att_5^{\uparrow}, Att_5^{\downarrow}, Att_5^{\leftarrow}, Att_5^{\rightarrow}, Att_5^{\nearrow}, Att_5^{\nwarrow}, Att_5^{\swarrow})$ , 其定义类似。取 PANet 的最后一个 RGB 特征映射为  $F_5$ 。用核大小为  $3 \times 3$  的卷积层对  $F_5$  进行卷积, 然后执行 Sigmoid 函数以获得特征图, 定义如公式(4)所示:

$$Att^{\searrow} = \sigma(W^{\searrow} * F_5) \quad (4)$$

其中  $W^{\searrow}$  是  $3 \times 3$  卷积的权值,  $\sigma$  表示激活函数 sigmoid。其次, 从特征图的周围像素中获取方向注意图, 计算如公式(5)所示:

$$\begin{cases} Att_5^{\searrow}[k]_{[i,j]} = (1 - Att^{\searrow}[k]_{[i,j]}) F_5[k]_{[i,j]} + \\ \quad Att^{\searrow}[k]_{[i,j]} F_5[k-1]_{[i-1,j-1]} \quad (j < H) \\ Att_5^{\searrow}[k]_{[j]} = (1 - Att^{\searrow}[k]_{[j]}) F_5[k]_{[j]} + \\ \quad Att^{\searrow}[k]_{[j]} F_5[k-1]_{[j-1]} \quad (j \geq H) \end{cases} \quad (5)$$

其中  $(i, j)$ ,  $[k]_{[i,j]}$  和  $H$  分别表示特征图的坐标、位置和高度。由于考虑了  $\searrow$  方向的注意模式,  $[k-1]_{[i-1,j-1]}$  和  $[k-1]_{[j]}$  分别代表  $[k]_{[i,j]}$  的左上角和上方。该操作可以使当前位置的值与其相邻位置的信息相关联。因此, 当篡改区域被修复时, 八方向局部注意模块可以很好地聚焦于该区域及其邻近区域, 以检测这些不一致的相关性。

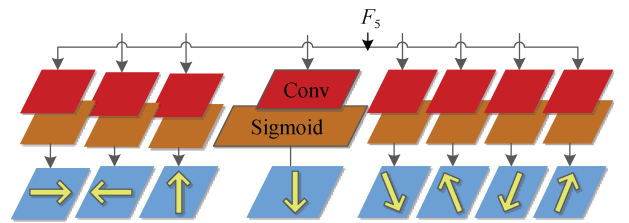


图 5 八方向局部注意模块

Figure 5 Eight-direction local attention module

### 3.5 级联 ConvGRU 的解码器模块

虽然深度视频修复的目标是产生具有时间一致性的更真实的逼真内容, 但由于缺乏处理运动遮挡以及非刚性运动目标的能力, 仍然存在一些时间不一致性。因此, 相邻帧之间缺乏时间一致性就成为篡改区域定位的重要线索。本文在解码器中引入八个 ConvGRU 的级联来捕捉时空不一致性。特别是, 为了防止随着网络的深化, 篡改区域中难以察觉的空间不一致痕迹消失, 在编码模块中使用不同尺度的

特征图和一个方向局部注意力图作为 ConvGRU 的每个 ConvGRU 单元的输入, 捕捉时间不一致, 然后产生像素级的预测。

解码器模块包含八个级联 ConvGRU, 每个级联 ConvGRU 包含 4 个 Cell 和 4 个上采样, 以处理对应于编码器模块的 4 个不同尺度的特征地图。具体地说, 8 个级联 ConvGRU 对应于八方向局部注意模块的 8 个注意映射。每个单元的核大小设置为  $3 \times 3$ , 通过 4 次  $2 \times$  上采样将输出的空间分辨率放大到与输入帧相同的大小, 以降低训练难度。同时, 对每个 Cell 进行双线性核上采样, 使每个 Cell 的隐含单元的输出与多模态特征图的大小相对应, 以便与不同尺度的多模态特征图进行融合。此外, 每个 Cell 计算出的隐藏状态为下一个 Cell 提供输入。形式上, 对于第  $t$  个时间步长, 第  $i$  ( $2 \leq i \leq 4$ ) 个 Cell 计算第  $t+1$  个时间步的隐藏状态, 计算如公式(6)所示:

$$\begin{aligned} h_{i+1} &= \text{ConvGRU}_i(F_i, h_i) \quad (2 \leq i \leq 4) \\ F_i^t &= [U(h_{i-1}^t), F_{6-i}^t] \end{aligned} \quad (6)$$

其中,  $h_i$ ,  $F_i$  和  $U$  分别为第  $i$  个 Cell 的隐藏状态、多模式特征映射的第  $i$  个尺度和上采样;  $F_i^t$ ,  $h_i^t$  和  $F_{6-i}^t$  分别表示  $t$  个时间步长的第  $i$  个多模式特征、第  $i$  个 Cell 的隐藏状态和第  $6-i$  个多模式特征。为了处理上采样产生的棋盘状伪影, 使用一个额外的  $3 \times 3$  卷积来减少这个问题。最后, 使用 Sigmoid 层进行像素级预测, 得到定位结果。

### 3.6 损失函数

在本文中, 选择 IoU 损失作为像素级预测。IoU 损失可以将候选对象的边界框作为一个整体进行回归, 不仅收敛速度快, 并且目标定位也更加准确, 计算如公式(7)所示:

$$L(o, g) = 1 - \frac{\sum O * G}{\sum (O + G - O * G) + \varepsilon} \quad (7)$$

其中  $O$ ,  $G$ ,  $\varepsilon$  和  $*$  分别表示输出结果、真实掩膜、一个能避免零除的小数字以及交集。当级联 ConvGRU 执行候选视频来捕获时空信息时, 损失会被更新。因此, 通过对时空信息的反复回归, 使得篡改区域的定位变得更加准确。

### 3.7 实现细节

基于级联 ConvGRU 和八方向局部注意力的深度视频修复篡改的被动取证采用编、解码器架构, 各层的输入输出尺寸、卷积核等参数如表 1 所示。

## 4 实验及分析

在本节中, 使用该方法与深度视频修复检测

方法 VIDNet<sup>[5]</sup>和深度图像修复检测方法 HPF<sup>[31]</sup>进行了比较, 且所有实验结果都是采用公开的源码测试所得, 同时还对该方法进行了各种泛化实验与消融实验以及在噪声和压缩状态下的性能分析。

表 1 网络模型参数

Table 1 The parameters of the proposed network					
模块	层数	核尺寸	核数目	步长	输出尺寸
编码器	Conv1	$7 \times 7$	64	1	$m \times n$
	Layer1	$3 \times 3$	256	2	$m/2 \times n/2$
	Layer2	$3 \times 3$	512	2	$m/4 \times n/4$
	Layer3	$3 \times 3$	1024	2	$m/8 \times n/8$
	Layer4	$3 \times 3$	2048	2	$m/16 \times n/16$
	Conv2	$3 \times 3$	256	1	$m/8 \times n/8$
	Conv3	$3 \times 3$	256	1	$m/4 \times n/4$
	Conv4	$3 \times 3$	256	1	$m/2 \times n/2$
八方向	Conv5	$3 \times 3$	256	1	$m/16 \times n/16$
解码器	Cell1	$3 \times 3$	256	1	$m/16 \times n/16$
	UP	—	—	—	$m/8 \times n/8$
	Cell2	$3 \times 3$	256	1	$m/8 \times n/8$
	UP	—	—	—	$m/4 \times n/4$
	Cell3	$3 \times 3$	256	1	$m/4 \times n/4$
	UP	—	—	—	$m/2 \times n/2$
	Cell4	$3 \times 3$	256	1	$m/2 \times n/2$
	UP	—	—	—	$m \times n$
	Conv6	$3 \times 3$	64	1	$m \times n$
	Conv7	$3 \times 3$	1	1	$m \times n$

为了更好的显示本文所提模型的性能, 在同等环境下, 将其与 VIDNet 在参数、运行速度上进行了对比, 如表 2 所示。由表 2 可知, 本文所提方法, 运行速度更快, 参数量更少, 更轻量化。

表 2 网络模型对比表

Table 2 The comparison table of network model			
方法	编码器参数	解码器参数	运行时间(迭代一次)
VIDNet	285342736	52619951	9/s
本文	49126416	31418931	4.5/s

### 4.1 数据集和实验细节

该模型基于 Pytorch 框架实现, 在 NVIDIA GeForce RTX 2080Ti 上训练与测试。在实验中, 使用 Davis 2016 作为修复数据集, 包含 50 个视频, 其中, 30 个用于训练, 20 个用于测试。网络的输入尺寸为  $240 \times 427$ , 每次将连续的 3 帧送入网络。该模型使用 Adam 优化器, 编码器的学习率固定为  $1 \times 10^{-6}$ , 解码器的学习率固定为  $1 \times 10^{-3}$ , 权重衰减固定为  $1 \times 10^{-6}$ 。训练数据使用了垂直翻转和水平翻转来增广数据集,

迭代次数为 40。其次, 本文使用 FGVC<sup>[4]</sup>, DSTT<sup>[3]</sup>, VI<sup>[2]</sup>, OPN<sup>[1]</sup>这 4 种深度修复方法合成修复视频, 并使用 IoU 损失作为评估指标。

4.2 实验结果

为了显示该网络结构的检测能力和泛化效果, 进行了类间测试与类内测试。在类内测试中, 按照 Davis 的默认设置, 在同种修复方法中, 30 个用于训练, 20 个用于测试, 并使用数据增广。在类间测试中, 选择 VI 修复数据集用于训练, 其他的 OP、DSTT、FGVC 修复的视频集用于测试。

4.2.1 类内测试

类内测试以及对比结果如表 3 和图 6 所示, 本文提出方法在四种修复结果上的定位效果如图 7-图 10 所示。由实验结果可知, 由于 VI 和 OP 所采用的视频修复模型并没有很好的考虑到空间特征, 因此修复后的视频具有很强的边缘效应。因此, 在进行特征提取时, 能较为容易的提取到篡改区域的特征, 故而 VI 和 OP 方法的篡改视频具有很好的定位效果,

能检测出更多的细节; DSTT 方法次之, 但是依然能较好地定位修复区域; FGVC 方法合成的篡改视频, 由于修复过程采用空时 transformer, 修复后的视频更自然流畅, 因此检测效果较差; 同时, 在对比实验中, PANet 作为特征提取骨干在提取深层特征时能很好的解决特征退化问题和梯度爆炸问题; 而八方向局部注意力模块从 8 个方向来关注当前像素的相邻信息, 使得提取到的特征具有更多的语义信息, 且检测器模块考虑到了时域特性。因此, 本文所提方法优于 HPF, 性能接近 VIDNet, 能精准定位篡改区域。

表 3 类内测试 IoU 值

Table 3 IoU values of the intra-class test				
Methods	VI*	OP*	DSTT*	FGVC*
HPF	0.46	0.49	0.38	0.27
VIDNet	0.60	0.61	0.49	0.40
本文	0.56	0.57	0.46	0.37

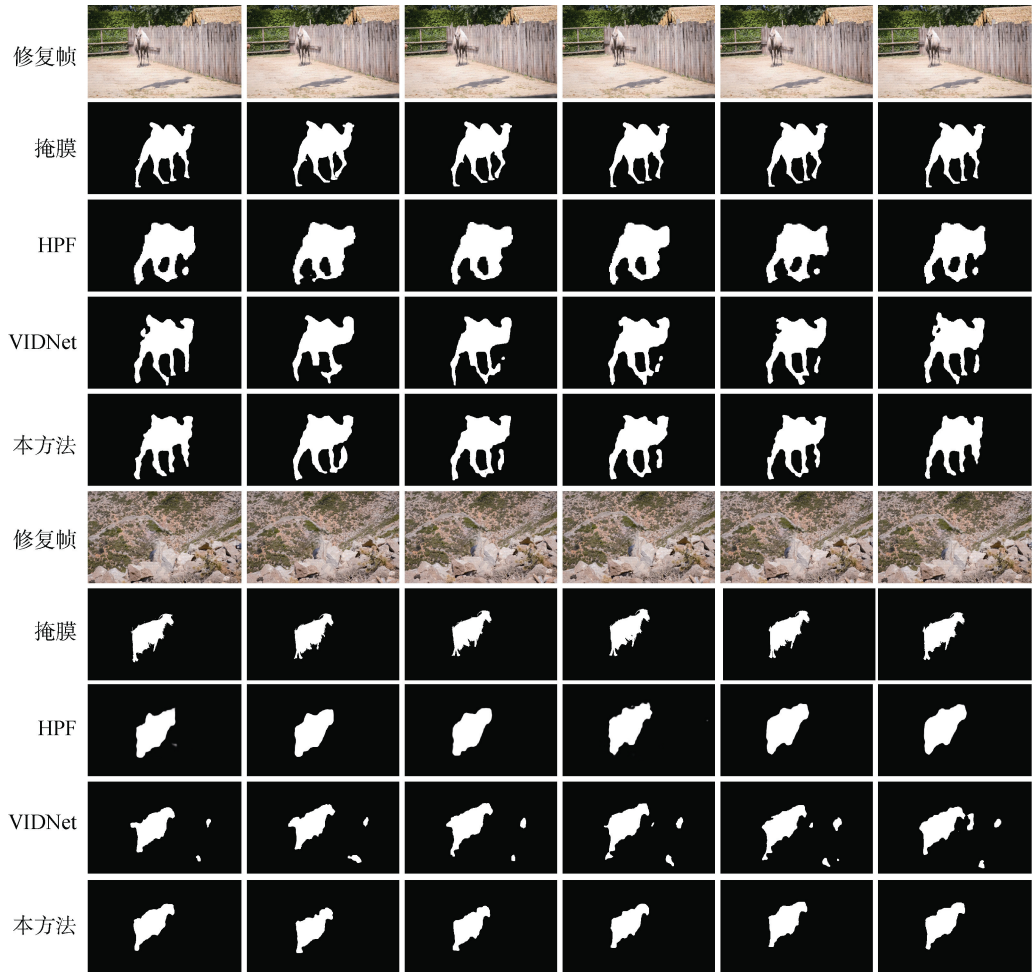


图 6 类内对比实验的定位效果

Figure 6 The localization results of the comparative experiment in the intra-class test



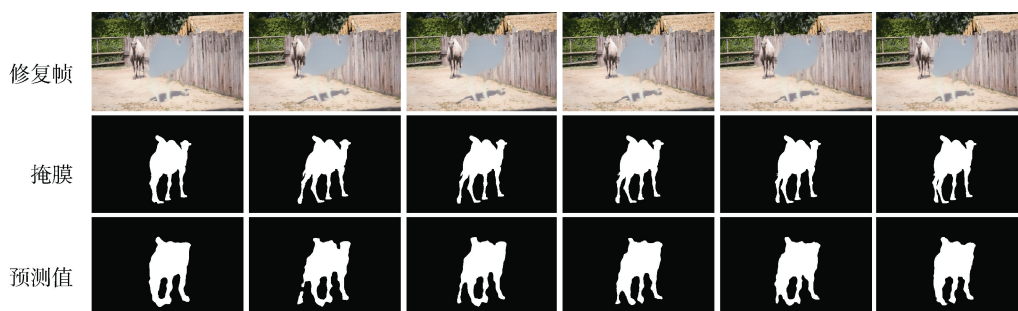


图 7 本文提出方法在 VI 修复视频上定位效果

Figure 7 The localization results of the proposed method for VI method



图 8 本文提出方法在 OP 修复视频上的定位效果

Figure 8 The localization results of the proposed method for OP method



图 9 本文提出方法在 DSTT 修复视频上的定位效果

Figure 9 The localization results of the proposed method for DSTT method



图 10 本文提出方法在 FGVC 修复视频上的定位效果

Figure 10 The localization results of the proposed method for FGVC method

#### 4.2.2 类间测试

在类间测试中, 使用 VI 做训练, OP、DSTT 和 FGVC 做测试。类间测试及对比结果如表 4 和图 11 所示。由实验结果可知, 该模型对于篡改区域规则且

篡改区域较大的篡改视频具有较好的定位效果。但当篡改区域形状不规则, 篡改细节较小时, 其定位效果较差。虽能定位到大概的篡改位置, 但预测出来的篡改区域会出现信息丢失现象, 且检测不到细节。归其

原因, 不同的视频修复算法在对视频进行修复时, 其修复效果不同, 遗留的修复痕迹各异。如 VI、OP 的视频修复效果较差, 能看出修复痕迹, DSTT 和 FGVC 修复效果好, 看起来自然流畅。因此当取 VI 数据集做训练时, 由于网络模型只学习到 VI 数据集的特征, 当取 OP、DSTT 和 FGVC 做测试时, 定位效果较差。

表 4 类间测试 IoU 值

Table 4 Inter-class test IoU values

Methods	OP	DSTT	FGVC
HPF	0.21	0.26	0.16
VIDNet	0.27	0.25	0.15
本文	0.26	0.24	0.14

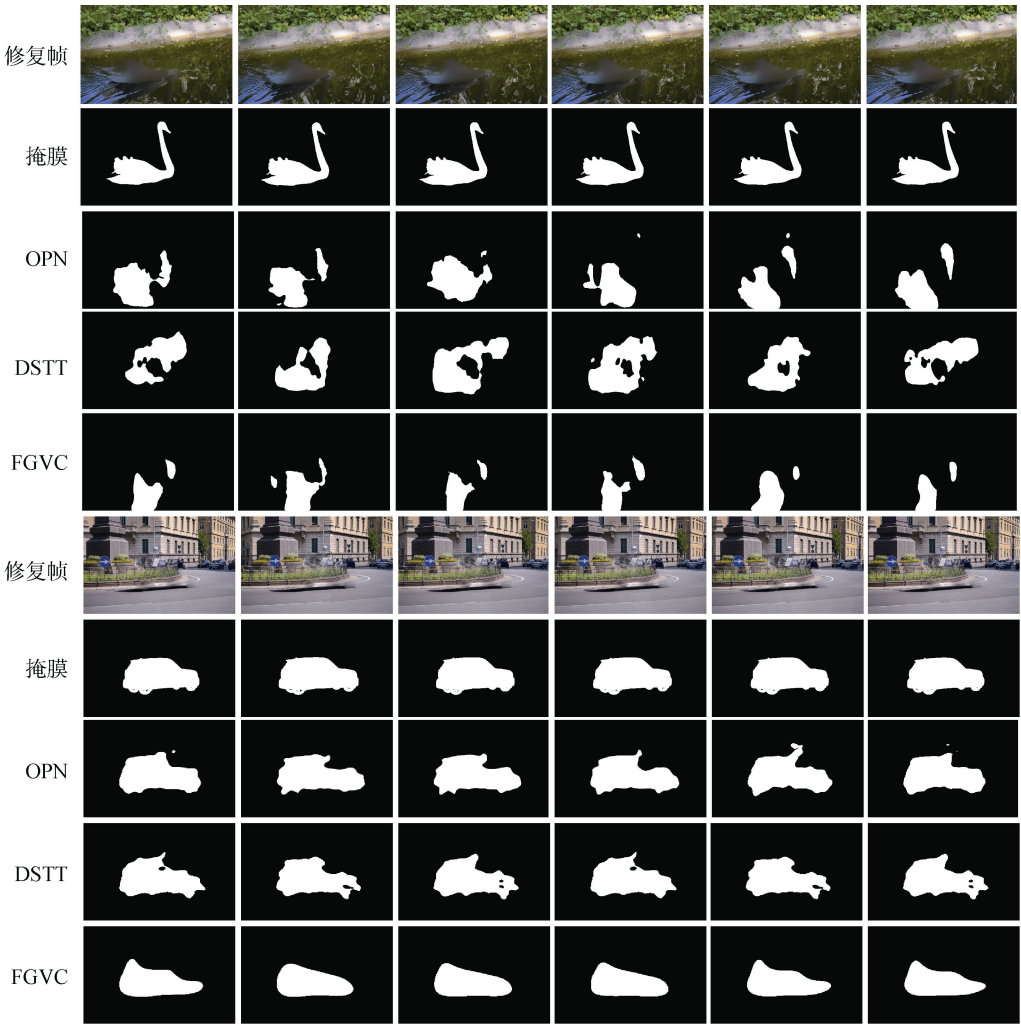


图 11 类间对比实验的定位效果图

Figure 11 The localization results of the comparative experiment in the inter-class test

由以上实验结果可知, 生成修复视频的方法越好, 检测效果越差, 且所有方法与的性能都较差(VI 做训练, OP 做测试, 对比结果如图 12 所示), 可知模型的泛化性能上还有待提升。

4.3 消融实验

为了分析本文所提方法模型的各个组件的有效性, 使用 VI 合成的深度修复数据集进行训练和测试, 消融实验设计如下:

- 不同的输入视频帧: 2 帧、4 帧;
- 单个输入流: 仅 RGB 或仅 ELA;
- VGG 代替 PANet 作为骨干网络;

- 对骨干网络中的卷积采用空间可分离卷积和空洞卷积替换;
- 将特征级联融合替换为特征相加融合或特征平均融合;
- 不同的 ConvGRU 的 Cell 层数: 3 层、5 层;
- 使用 BN 层替换 IN 层;
- 将 ELA 替换为噪声 SRM;
- 八方向局部注意模块的变化: 四方向局部注意模块替换或去掉改组件;
- 其他注意机制替换八方向注意力模块: 残差通道注意力(Residual Groups With Channel Atten-

tion, RCA)<sup>[37]</sup>、卷积块注意力模块(Convolutional Block Attention Module, CBAM)<sup>[38]</sup>或通道注意力机制(Efficient Channel Attention, ECA)<sup>[39]</sup>。

• 将八方向注意力模块分别添加到不同特征层中: 第四层特征(f4)、第三层特征(f3)、第二层特征(f2)。

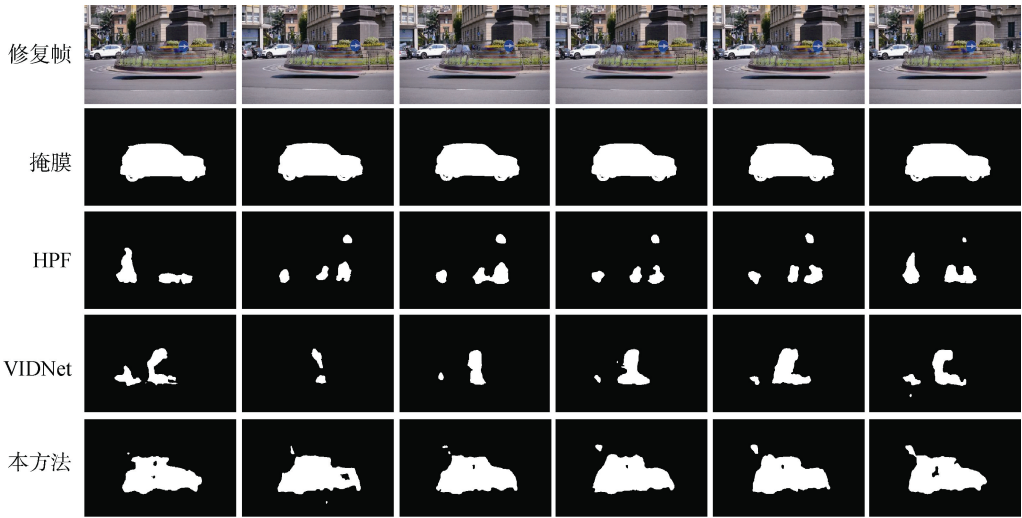


图 12 不同检测方法类间定位效果图

Figure 12 Inter-class positioning effect diagram of different detection methods

表 5 消融实验结果			
Table 5 Results of ablation experiments			
变体	IoU	变体	IoU
2 帧	0.56	BN	0.56
4 帧	0.57	RGB 与 SRM	0.56
仅 RGB 帧	0.54	四方向	0.57
仅 ELA 帧	0.48	去掉八方向	0.56
VGG	0.59	RCA	0.53
空间可分离卷积	0.56	CBAM	0.54
空洞卷积	0.56	ECA	0.53
5 层 Cell	0.54	f4+八方向	0.55
3 层 Cell	0.54	f3+八方向	0.55
特征相加融合	0.56	f2+八方向	0.55
特征平均融合	0.55	本文方法	0.58

注: 使用 VI 做训练, VI 做测试。

消融实验中,除了替换模块,其他模块和参数保持不变,各种变体的实验结果如表 5 所示。从表中可以得出以下结论: 3 帧足以保持视频帧间的连续性; 双流网络使得提取到的语义信息更为丰富; ELA 特征与 RGB 特征的融合,使得提取到的特征更为详细; PANet 中增加的自底向上路径能很好的传递浅层的定位信息; 4 层 Cell 刚好对应 4 个特征,起到适配作用; 对于小 Batch, IN 的性能优于 BN; 八方向局部注意力模块能更好的关注像素间的关系,使得提取到的信息更为平滑,语义信息更为丰富。因此,可以得出结论,ELA、RGB、PANet、八方向局部注意力模块、4 个 ConvGRU 的组合属于所有变体里的最优组

合。然而,通过对消融实验的分析,模型的检测精度提升空间并不是很大,未来是否应该考虑更好的检测模型来提升篡改对象细节信息的检测,使得模型的检测精度更为精准。

4.4 对不同视频压缩的鲁棒性实验

为了测试本文所提方法在视频被压缩下的鲁棒性,对 VI 合成的修复视频和原始视频,用 HEVC 和 H.264 分别对视频进行 1M/s、2M/s 和 3M/s 的压缩,并与 HPF 和 VIDNet 对比,实验结果如表 6 所示。

由表 6 可看出,视频压缩比特率越低,压缩越严重,所有方法的 IoU 值都出现了不同程度的降低。且与 HPF 相比,略显优势,性能接近 VIDNet,因为本文所提方法利用八方向局部注意力机制和级联的 ConvGRU 考虑了更多的伪造时空线索。

4.5 对不同视频添加噪声的健壮性实验

为了测试本文所提方法在噪声干扰下的健壮性,对现有的数据集添加不同级别的随机高斯噪声,即 30db、40db 和 50db,并与 HPF 和 VIDNet 对比,实验结果如表 7 所示。

从实验结果中可知,当添加随机噪声,所有模型的性能都存在降低,且信噪比越低, IoU 越低。

5 结论

本文提出了一种端到端的视频修复定位取证模型。该方法主要针对深度视频修复过程中留下的时空痕迹,由编码器和解码器模块捕获。需要强调的是,

表 6 HEVC 和 H.264 压缩下的实验结果

Table 6 Experimental results for HEVC and H.264 compression

修复方法		HPF	VIDNet	本文
VI*	HEVC	1M/s 0.43	0.49	0.47
		2M/s 0.44	0.54	0.52
		3M/s 0.45	0.56	0.53
	H.264	1M/s 0.41	0.51	0.49
		2M/s 0.43	0.57	0.55
		3M/s 0.44	0.58	0.56
OP*	HEVC	1M/s 0.37	0.43	0.40
		2M/s 0.42	0.48	0.46
		3M/s 0.45	0.52	0.51
	H.264	1M/s 0.32	0.38	0.35
		2M/s 0.39	0.44	0.42
		3M/s 0.42	0.48	0.46
DSTT*	HEVC	1M/s 0.32	0.43	0.40
		2M/s 0.35	0.46	0.42
		3M/s 0.36	0.48	0.43
	H.264	1M/s 0.31	0.41	0.38
		2M/s 0.33	0.44	0.41
		3M/s 0.35	0.47	0.44
FGVC*	HEVC	1M/s 0.22	0.35	0.33
		2M/s 0.24	0.37	0.34
		3M/s 0.25	0.38	0.35
	H.264	1M/s 0.21	0.31	0.29
		2M/s 0.23	0.35	0.30
		3M/s 0.24	0.39	0.35

注: 带“\*”表示训练。

表 7 噪声实验结果

Table 7 Noise test results

方法	HPF	VIDNet	本文
VI*	30db 0.40	0.49	0.47
	40db 0.39	0.48	0.45
	50db 0.37	0.46	0.45
OP*	30db 0.24	0.30	0.28
	40db 0.23	0.31	0.30
	50db 0.22	0.29	0.27
DSTT*	30db 0.29	0.39	0.36
	40db 0.28	0.38	0.37
	50db 0.26	0.36	0.35
FGVC*	30db 0.20	0.29	0.27
	40db 0.18	0.27	0.25
	50db 0.17	0.28	0.26

注: 带“\*”表示训练。

引入了八方向局部注意模块和级联 ConvGRU, 以便更好地表示修复区域的时空不一致性。连续 RGB 帧和 ELA 两种类型的模态被发送到编码器模块的两个

并行 PANet 中用于学习判别特征。为了进一步利用时间相关性, 在解码器模块中嵌入 8 个具有 4 个单元的级联 ConvGRU。与已有的方法相比, 本文所提方法定位精准, 且运行速度更快、参数量更少, 模型更轻量化。

致谢 在此向给本文给予指导的老师, 以及提供帮助的同学和给本文提出建议的审稿专家表示诚挚的感谢。

参考文献

[1] Oh S W, Lee S, Lee J Y, et al. Onion-Peel Networks for Deep Video Completion[C]. 2019 IEEE/CVF International Conference on Computer Vision, 2019: 4402-4411.

[2] Kim D, Woo S, Lee J Y, et al. Deep Video Inpainting[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 5785-5794.

[3] Liu R, Deng H M, Huang Y Y, et al. Decoupled Spatial-Temporal Transformer for Video Inpainting[EB/OL]. 2021: arXiv: 2104.06637. <http://arxiv.org/abs/2104.06637>.

[4] Gao C, Saraf A, Huang J B, et al. Flow-Edge Guided Video Completion[C]. European Conference on Computer Vision, 2020: 713-729.

[5] Zhou P, Yu N, Wu Z X, et al. Deep Video Inpainting Detection[EB/OL]. 2021: arXiv: 2101.11080. <http://arxiv.org/abs/2101.11080>.

[6] Ding X L, Yang G B, Li R, et al. Identification of Motion-Compensated Frame Rate Up-Conversion Based on Residual Signals[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(7): 1497-1512.

[7] Ding X L, Pan Y F, Luo K, et al. Localization of Deep Video Inpainting Based on Spatiotemporal Convolution and Refinement Network[C]. 2021 IEEE International Symposium on Circuits and Systems, 2021: 1-5.

[8] Liu G L, Reda F A, Shih K J, et al. Image Inpainting for Irregular Holes Using Partial Convolutions[C]. Computer Vision – ECCV 2018: 15th European Conference, 2018: 89-105.

[9] Pathak D, Krähenbühl P, Donahue J, et al. Context Encoders: Feature Learning by Inpainting[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2536-2544.

[10] Yu J H, Lin Z, Yang J M, et al. Free-Form Image Inpainting with Gated Convolution[C]. 2019 IEEE/CVF International Conference on Computer Vision, 2019: 4470-4479.

[11] Yang C, Lu X, Lin Z, et al. High-Resolution Image Inpainting Using Multi-Scale Neural Patch Synthesis[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4076-4084.

[12] Barnes C, Shechtman E, Finkelstein A, et al. PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing[M]. Seminal Graphics Papers: Pushing the Boundaries, Volume 2. New York, NY, USA: ACM, 2023: 619-629.

[13] Huang J B, Kang S B, Ahuja N, et al. Temporally Coherent Completion of Dynamic Video[J]. ACM Transactions on Graphics, 35(6): 196.



- [14] Zeng Y H, Fu J L, Chao H Y. Learning Joint Spatial-Temporal Transformations for Video Inpainting[C]. *European Conference on Computer Vision*, 2020: 528-543.
- [15] Li Q, Wang R D, Xu D W. Detection to Video Moving Object Deletion Based on Video Inpainting[J]. *Journal of Optoelectronics-Laser*, 2016, 27(2): 182-190.  
(李倩, 王让定, 徐达文. 基于视频修复的运动目标删除篡改行为的检测算法[J]. *光电子·激光*, 2016, 27(2): 182-190.)
- [16] Bai S S, Ni R R, Zhao Y. Video Forensics for Object Removal Based on Two-Channel Convolutional Neural Network[J]. *Journal of Signal Processing*, 2020, 36(9): 1415-1421.  
(白珊山, 倪蓉蓉, 赵耀. 基于双通道卷积神经网络的视频目标移除取证算法[J]. *信号处理*, 2020, 36(9): 1415-1421.)
- [17] Xiong X, Huang Z, Xu C, et al. Digital Video Forgeries Detection Based on Prediction Error[J]. *Information Security and Communications Privacy*, 2008, 6(12): 128-130.  
(熊潇, 黄征, 徐彻, 等. 基于预测残差检测的数字视频篡改鉴定[J]. *信息安全与通信保密*, 2008, 6(12): 128-130.)
- [18] Huang T Q, Chen Z W, Su L C, et al. Digital Video Forgeries Detection Based on Content Continuity[J]. *Journal of Nanjing University (Natural Sciences)*, 2011, 47(5): 493-503.  
(黄添强, 陈智文, 苏立超, 等. 利用内容连续性的数字视频篡改检测[J]. *南京大学学报(自然科学版)*, 2011, 47(5): 493-503.)
- [19] Amerini I, Galteri L, Caldelli R, et al. Deepfake Video Detection through Optical Flow Based CNN[C]. *2019 IEEE/CVF International Conference on Computer Vision Workshop*, 2019: 1205-1207.
- [20] Güera D, Delp E J. Deepfake Video Detection Using Recurrent Neural Networks[C]. *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2018: 1-6.
- [21] Saxena S, Subramanyam A V, Ravi H. Video Inpainting Detection and Localization Using Inconsistencies in Optical Flow[C]. *2016 IEEE Region 10 Conference*, 2016: 1361-1365.
- [22] Zhu Y, Chen C F, Yan G, et al. AR-Net: Adaptive Attention and Residual Refinement Network for Copy-Move Forgery Detection[J]. *IEEE Transactions on Industrial Informatics*, 2020, 16(10): 6714-6723.
- [23] Hsu C C, Hung T Y, Lin C W, et al. Video Forgery Detection Using Correlation of Noise Residue[C]. *2008 IEEE 10th Workshop on Multimedia Signal Processing*, 2008: 170-174.
- [24] Lin C S, Tsay J J. A Passive Approach for Effective Detection and Localization of Region-Level Video Forgery with Spatio-Temporal Coherence Analysis[J]. *Digital Investigation*, 2014, 11(2): 120-140.
- [25] Aminu Bagiwa M, Abdul Wahab A W, Idna Idris M Y, et al. Digital Video Inpainting Detection Using Correlation of Hessian Matrix[J]. *Malaysian Journal of Computer Science*, 2016, 29(3): 179-195.
- [26] Yuan X J, Huang T Q, Su L C, et al. Detection of Video Matting Tamper Based on Edge Anomaly and Compressive Tracking[J]. *Computer Engineering*, 2014, 40(7): 267-271, 276.  
(袁秀娟, 黄添强, 苏立超, 等. 基于边缘异常与压缩跟踪的视频抠像篡改检测[J]. *计算机工程*, 2014, 40(7): 267-271, 276.)
- [27] Zhang J, Chen J, Su Y T. Detection of Region-Duplication Forgery in the Video Streams[J]. *Electronic Measurement Technology*, 2011, 34(11): 66-69.  
(张静, 陈静, 苏育挺. 基于滤波检测的视频区域篡改检测算法[J]. *电子测量技术*, 2011, 34(11): 66-69.)
- [28] Weng S W, Peng Y H, Wei B, et al. A Two-Stage Algorithm for Video Forgery Detection Based on Inception-V3 Network[J]. *Journal of Guangdong University of Technology*, 2019, 36(6): 16-23.  
(翁韶伟, 彭一航, 危博, 等. 基于 Inception-V3 网络的双阶段数字视频篡改检测算法[J]. *广东工业大学学报*, 2019, 36(6): 16-23.)
- [29] Chen L Q, Yang Q X, Yuan L F, et al. Passive Forensic Based on Spatio-Temporal Localization of Video Object Removal Tampering[J]. *Journal on Communications*, 2020, 41(7): 110-120.  
(陈临强, 杨全鑫, 袁理锋, 等. 视频对象移除篡改的时空域定位被动取证[J]. *通信学报*, 2020, 41(7): 110-120.)
- [30] Li A, Ke Q H, Ma X J, et al. Noise Doesn't Lie: Towards Universal Detection of Deep Inpainting[EB/OL]. 2021: arXiv: 2106.01532. <http://arxiv.org/abs/2106.01532>.
- [31] Li H D, Huang J W. Localization of Deep Inpainting Using High-Pass Fully Convolutional Network[C]. *2019 IEEE/CVF International Conference on Computer Vision*, 2019: 8300-8309.
- [32] Ding X L, Pan Y F, Luo K, et al. Localization of Deep Video Inpainting Based on Spatiotemporal Convolution and Refinement Network[C]. *2021 IEEE International Symposium on Circuits and Systems*, 2021: 1-5.
- [33] Yu B Y, Li W H, Li X, et al. Frequency-Aware Spatiotemporal Transformers for Video Inpainting Detection[C]. *2021 IEEE/CVF International Conference on Computer Vision*, 2021: 8168-8177.
- [34] Perazzi F, Pont-Tuset J, McWilliams B, et al. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 724-732.
- [35] Lin T Y, Dollár P, Girshick R, et al. Feature Pyramid Networks for Object Detection[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 936-944.
- [36] Liu S, Qi L, Qin H F, et al. Path Aggregation Network for Instance Segmentation[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 8759-8768.
- [37] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you Need[C]. *The 31st International Conference on Neural Information Processing Systems*, 2017: 6000-6010.
- [38] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional Block Attention Module[C]. *European Conference on Computer Vision*, 2018: 3-19.
- [39] Wang Q L, Wu B G, Zhu P F, et al. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 11531-11539.



**熊义毛** 于 2020 年在湖南科技大学潇湘学院计算机科学与技术专业获得学士学位, 现在湖南科技大学计算机科学与工程学院电子信息专业攻读硕士学位。CCF 会员, 研究领域为多媒体内容安全。E-mail: 1203661414@qq.com



**丁湘陵** 于 2018 年在湖南大学计算机科学与技术专业获得博士学位, 现任湖南科技大学教授、硕士生导师, 研究领域为多媒体内容安全、图像/视频处理等。E-mail: xianglingding@hnust.edu.cn



**谷庆** 于 2019 年在安徽新华学院学院软件工程专业获得学士学位, 现在湖南科技大学计算机科学与工程学院电子信息专业攻读硕士学位。研究领域为多媒体内容安全。E-mail: 3334143570@qq.com



**杨高波** 于 2004 年在上海大学通信与信息系统专业获得博士学位, 现任湖南大学教授、博士生导师, 研究领域为图像/视频信号处理和多媒体通信, 包括图像/视频信息安全, 包括图像/视频篡改取证, 压缩域视频水印, 高效视频编码及其优化与大数据权属保护技术。E-mail: yanggaobo@hnu.edu.cn



**赵险峰** 于 2003 年在上海交通大学计算机科学与工程系获得博士学位, 现任中国科学院信息工程研究所研究员、中国科学院大学教授、IJDCF、IWDW 等期刊、会议的编委、主席或委员, 任中国电子学会通信与信息安全专委会、中国图象图形学会多媒体取证与安全专委会等学术组织的委员, 研究领域为多媒体安全与智能分析的理论与技术, 包括多媒体处理与智能分析、多媒体信息隐藏与检测、多媒体伪造取证与防护、多媒体内容智能生成、数字水印与多媒体版权保护等。E-mail: zhaoxianfeng@iie.ac.cn