软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(5):2051−2067 [doi: 10.13328/j.cnki.jos.006768]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

预训练驱动的多模态边界感知视觉 Transformer*

石泽男[1,2], 陈海鹏[1,2], 张 冬[3], 申铉京[1,2]

[1](吉林大学 计算机科学与技术学院, 吉林 长春 130012)

[2](符号计算与知识工程教育部重点实验室 (吉林大学), 吉林 长春 130012)

[3](香港科技大学 计算机科学与工程系, 香港 999077)

通信作者: 陈海鹏, E-mail: chenhp@jlu.edu.cn

摘 要: 卷积神经网络 (convolutional neural network, CNN) 在图像篡改检测任务中不断取得性能突破, 但在面向真实场景下篡改手段未知的情况时, 现有方法仍然无法有效地捕获输入图像的长远依赖关系以缓解识别偏差问题, 从而影响检测精度. 此外, 由于标注困难, 图像篡改检测任务通常缺乏精准的像素级图像标注信息. 针对以上问题, 提出一种预训练驱动的多模态边界感知视觉 Transformer. 首先, 为捕获在 RGB 域中不可见的细微伪造痕迹, 引入图像的频域模态并将其与 RGB 空间域结合作为多模态嵌入形式. 其次利用 ImageNet 对主干网络的编码器进行训练以缓解当前训练样本不足的问题. 然后, Transformer 模块被整合到该编码器的尾部, 以达到同时捕获低级空间细节信息和全局上下文的目的, 从而提升模型的整体表征能力. 最后, 为有效地缓解因伪造区域边界模糊导致的定位难问题, 构建边界感知模块, 其可以通过 Scharr 卷积层获得的噪声分布以更多地关注噪声信息而不是语义内容, 并利用边界残差块锐化边界信息, 从而提升模型的边界分割性能. 大量实验结果表明, 所提方法在识别精度上优于现有的图像篡改检测方法, 并对不同的篡改手段具有较好的泛化性和鲁棒性.

关键词: 模型预训练; 多模态; 视觉 Transformer; 边界感知; 图像篡改检测

中图法分类号: TP391

中文引用格式: 石泽男, 陈海鹏, 张冬, 申铉京. 预训练驱动的多模态边界感知视觉Transformer. 软件学报, 2023, 34(5): 2051–2067. http://www.jos.org.cn/1000-9825/6768.htm

英文引用格式: Shi ZN, Chen HP, Zhang D, Shen XJ. Pre-training-driven Multimodal Boundary-aware Vision Transformer. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2051–2067 (in Chinese). http://www.jos.org.cn/1000-9825/6768.htm

Pre-training-driven Multimodal Boundary-aware Vision Transformer

SHI Ze-Nan[1,2], CHEN Hai-Peng[1,2], ZHANG Dong[3], SHEN Xuan-Jing[1,2]

[1](College of Computer Science and Technology, Jilin University, Changchun 130012, China)

[2](Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education (Jilin University), Changchun 130012, China)

[3](Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong 999077, China)

Abstract: Convolutional neural networks (CNN) have continuously achieved performance breakthroughs in image forgery detection, but when faced with realistic scenarios where the means of tampering is unknown, the existing methods are still unable to effectively capture the long-term dependencies of the input image to alleviate the recognition bias problem, which affects the detection accuracy. In addition, due to the difficulty in labeling, image forgery detection usually lacks accurate pixel-level image labeling information. Considering the above problems, this study proposes a pre-training-driven multimodal boundary-aware vision transformer. To capture the subtle forgery

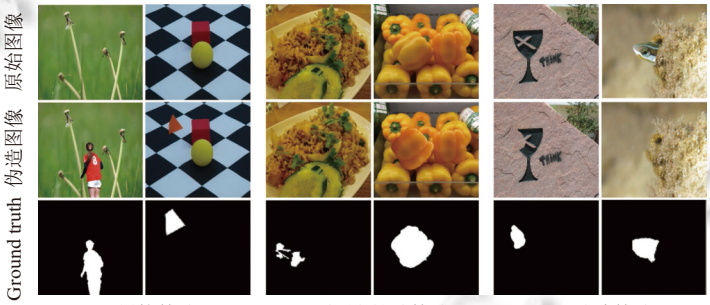
* 基金项目: 国家重点研发计划 (2018YFB0804202, 2018YFB0804203); 国家自然科学基金 (U19A2057, 61876070); 吉林大学 2021 年度“学科交叉融合创新”青年学者自由探索类项目 (JLUXKJC2021QZ01)
本文由“融合预训练技术的多模态学习研究”专题特约编辑宋雪萌副教授、聂礼强教授、申恒涛教授、田奇教授、黄华教授推荐.
收稿时间: 2022-04-15; 修改时间: 2022-05-29; 采用时间: 2022-08-24; jos 在线出版时间: 2022-09-20
CNKI 网络首发时间: 2023-03-17

traces invisible in the RGB domain, the method first introduces the frequency-domain modality of the image and combines it with the RGB spatial domain as a form of multimodal embedding. Secondly, the encoder of the backbone network is trained with ImageNet to alleviate the current problem of insufficient training samples. Then, the transformer module is integrated into the tail of this encoder to capture both low-level spatial details and global contexts, which improves the overall representation ability of the model. Finally, to effectively alleviate the problem of difficult localization caused by the blurred boundary of the forged regions, this study establishes a boundary-aware module, which can use the noise distribution obtained by the Scharr convolutional layer to pay more attention to the noise information rather than the semantic content and utilize the boundary residual block to sharpen the boundary information. In this way, the boundary segmentation performance of the model can be enhanced. The results of extensive experiments show that the proposed method outperforms existing image forgery detection methods in terms of recognition accuracy and has better generalization and robustness to different forgery methods.

Key words: model pre-training; multimodal; vision Transformer; boundary awareness; image forgery detection

视觉信号是人类感知外界信息最主要的途径之一. 近年来, 随着图像编辑和处理技术的发展, 人们借助 Photoshop、FakeApp 等工具可以轻松地获取、编辑图像的内容并以极低的成本生成篡改图像[1,2]. 然而, 图像篡改的盛行对我们的生活产生许多负面影响, 例如网络欺诈、虚假宣传与舆论操纵, 甚至学术造假[3]. 美国著名学术打假人伊丽莎白·比克 (Elisabeth Bik) 博士通过研究 20 621 篇论文发现其中 3.8% 的论文存在蓄意篡改图片的问题[4]. 此外, 据公开报道, 美国研究诚信办公室主任 John Dahlberg 表示: “图像篡改是一个日益显著的问题, 需要我们逐渐重视起来并进行解决.” 因此, 为了更好地保障社会秩序, 维护新闻诚信并保证网络内容安全, 及时充分地开发可靠的模型来揭示图像的篡改信息, 是信息安全领域迫切的现实需求.

早期的图像篡改检测方法主要集中在利用传统特征来判别图像是否被篡改的问题上, 仅有少数工作关注到像素级别的图像篡改检测[5]. 此外, 部分方法仅针对一种特定的图像篡改类型进行检测研究, 如图 1 所示的拼接篡改[6−9]、复制-粘贴篡改[10−12]和移除篡改[13]. 这些被精心篡改的图像肉眼看起来非常真实, 几乎与原始图像没有任何显著的视觉差异. 因此, 针对真实场景中更为复杂的图像篡改手段, 迫切需要新一代的算法, 以在像素级别上获得更精细的检测结果. 然而, 由于图像伪造区域存在变化尺度多样、形状不规则、边界模糊, 以及与真实区域相似度高等特点, 导致在像素层面上对图像篡改进行检测仍然面临较大的挑战.



(a) 拼接篡改 (b) 复制-粘贴篡改 (c) 移除篡改

图 1 不同篡改手段制造的伪造图像及对应的篡改区域示例图

近年来, 深度学习在计算机视觉等领域引领技术进步的潮流. 在此过程中, 国内外研究学者也逐渐将深度学习技术引入到图像篡改检测领域. 其中, MFCN[14]通过引入一个检测分支用来学习拼接区域的边界信息, 在图像篡改检测任务中展现出巨大的潜力, 有效地提升篡改区域的定位精度. 基于此思路, Bappy 等人[15]应用基于 LSTM 的图像块比较方法 (J-LSTM) 检测被篡改区域和真实区域的边界, 并进一步提出混合编码器-解码器结构 (H-LSTM)[16]来提高算法性能. 一些方法[17,18]在端到端框架之前利用 SRM (steganalysis rich model)[19]的 3 个高通滤波器来探索真实和被篡改区域之间的噪声不一致. 但由于 RGB-N[18]中采用 R-CNN 的网络结构, 导致其只能用矩形框标记篡改的区域. 然而, 在像素级别的图像篡改检测任务中, 上述方法在检测精度、特征泛化能力和鲁棒性方面距离实际应用还存在一定的差距. 基于此, 一些专家学者提出利用注意力模块来关注目标图像的重要区域. Zhu 等人[20]提出一种基于自适应注意力和残差细化的网络, 将位置和通道注意特征进行融合, 通过残差细

化模块对粗定位结果进行优化, 在图像拼接和复制-粘贴数据集上取得较好的检测效果. Hu 等人[21]提出一种基于空间金字塔注意力的网络 (SPAN), 通过构建基于局部自注意力的金字塔来有效地模拟多尺度图像块之间的关系, 从而提高检测精度.

上述方法虽然在图像篡改检测任务中取得良好的检测效果, 但仍存在以下两个问题: (1) 一方面, 基于 SRM 或注意力机制的编码-解码网络及其变体在特征提取过程中容易丢失部分全局上下文信息. 由于任何篡改行为都会在一定程度上破坏原始图像数据本身固有特征的完整性, 由此图像具有的一致性和独特性可作为自身的“固有指纹”用于鉴别伪造篡改. 因此, 对于图像篡改检测任务, 全局信息的提取是至关重要的[2,3]. 另一方面, 针对不同的篡改手段, 图像篡改检测任务中的模型在目标区域, 即篡改区域边界处的检测精度不够理想. (2) 上述先进的解决方案, 如 ManTra-Net[17]、RGB-N[18]和 SPAN[21], 均采用数据增强技术, 如图像翻转和图像旋转等操作, 以获得两倍或更多的训练样本, 避免模型过度拟合. 然而, 在这个领域中由于训练样本有限, 利用普通数据增强技术获得的额外样本数量仍然是有限的.

针对上述问题, 本文提出一种预训练驱动的多模态边界感知视觉 Transformer, 其能够精准地检测图像中伪造区域. 该模型主要包含以下 4 个模块: 频域模态、基于预训练的局部-全局特征增强模块、边界感知模块和渐进式语义生成模块. 本文首先基于 RGB 图像生成频域模态, 将多模态信息作为特征增强模块的输入, 其次利用模型预训练技术, 在不使用数据增强策略的前提下, 有效缓解图像编码器训练时样本不足的问题. 然后, 在主干网络的 CNN 编码器利用卷积层提取图像的高级语义特征后, 引入 Transformer 编码器进一步提取全局上下文信息. 最后, 将边界感知模块生成的特征图与伪造特征表示模块的输出作为输入送入渐进式语义生成模块, 逐步捕获空间和通道间的相关性, 引导网络关注目标区域, 从而提升篡改区域的检测精度. 本文的主要贡献如下.

(1) 在图像篡改检测任务中利用现有的图像分类数据集对主干编码网络进行预训练, 促进模型参数优化的同时又缓解训练数据不足的问题.

(2) 将多模态图像作为主干编码网络的输入, 并在编码器的尾部引入 Transformer 模块, 通过对 CNN 输出的高级语义特征进行再提取, 达到在空间上构建篡改图像全局上下文依赖关系的目的, 从而提升模型的表征能力.

(3) 为应对伪造区域边界模糊问题, 提出一种边界感知模块, 通过 Scharr 卷积层和边界残差块更多地关注图像噪声信息并捕捉篡改区域周围的边界伪影, 提升网络的边界分割性能.

(4) 实验结果表明, 本文提出的预训练驱动的多模态边界感知视觉 Transformer 在多个图像篡改数据集上取得的检测精度均优于基准模型和当前性能最优的方法, 并通过消融实验验证了本文方法的有效性.

1 相关工作

1.1 图像篡改检测

与传统方法相比, 基于深度学习的图像篡改检测方法对复杂数据具有更强的表征能力, 能够通过深度网络自动地提取具有判别能力的图像篡改特征. 目前, 基于深度学习的图像篡改检测方法主要包括噪声视图、边界监督和注意力机制方法等.

基于噪声视图的方法旨在利用拼接或移除篡改引入的新元素在噪声分布方面与真实部分存在不同的这一线索捕捉图像伪造痕迹, 以检测篡改区域. 针对一幅输入图像, 首先通过预先设定的高通滤波器或约束卷积层生成噪声视图, 然后以单独的[5,22,23]或与输入图像一起[17,18,21,24]的方式送入一个深度神经网络. 这些噪声流获得的噪声不一致性有助于增强图像的篡改痕迹, 然而该类方法对于检测没有引入新元素的复制-粘贴篡改是无效的.

基于边界监督的方法旨在增加一个辅助边界分支以捕捉被篡改区域周围的伪造痕迹. 其中北京邮电大学牛少彰团队应用基于 Sobel 边缘检测滤波器的 Mask R-CNN 检测篡改区域, 使预测的篡改掩码与真实掩码拥有相似的图像梯度[25]. MVSS-Net[24]也利用 Sobel 边缘检测滤波器构建一个边缘监督分支, 从而在伪造区域附近产生更集中的特征响应. GSR-Net[26]将来自不同层的骨干特征统一连接起来作为辅助分支的输入, 并利用一个判别生成器分割和细化图像篡改过程中产生的边界伪影. 然而可能存在一种风险, 即负责篡改检测的深层特征信息仍然是有限的, 导致在边界处和小目标区域的空间信息提取不足.

此外, 一些利用注意力模块来关注目标图像重要区域的图像篡改检测方法相继被提出, 其中, Islam 等人[12]利用基于双阶注意力的生成对抗网络 DOA-GAN 来检测和定位复制-粘贴篡改. Hu 等人[21]提出空间金字塔注意力网络 SPAN, 实现对拼接、复制-粘贴和移除 3 类图像篡改手段的像素级检测. Chen 等人[27]提出 RGB-频域注意力模块加强局部特征表达, 并利用多尺度相似模块来衡量局部特征之间的相似性, 同时实现换脸图像的篡改检测与定位. Liu 等人[28]通过注意力机制联合探索空间和通道的图像相关性和差异性, 获得更好的信息共享和更快的推理.

1.2 ImageNet 预训练

除模型本身的架构之外, 成功训练出优秀的网络模型的关键因素之一是对大规模数据集的良好利用. 与图像分类任务中的数据集规模相比, 其他图像处理任务的公开数据量相对较少, 如医学图像语义分割任务中的皮肤癌病灶数据集 ISIC 仅有 2 594 张图像, 图像复制-粘贴篡改数据集 COVERAGE 仅有 100 张伪造图像. 众所周知, 大规模数据集 ImageNet 中的图像多种多样, 且具有丰富的纹理和颜色信息. 随着深度学习在图像处理领域的广泛应用, 预训练已逐渐成为一种比较常规的策略, 如 GFFD[29]、MVSS-Net[24]和 PSCC-Net[28]模型中利用 ImageNet 上训练好的参数对它们的骨干网络 ResNet50、HRNetV2p-W18 和 Xception 进行参数初始化, 从而较好地训练所提出的复杂网络. 现阶段常见的 CNN 编码器如 ResNet、DenseNet 和 Inception 等网络结构层数很深, 包含几百万上千万的参数. 因此, 本文利用在 ImageNet 上预先训练好的 ResNet50 参数直接用于初始化主干网络结构, 促进优化模型参数的同时又解决训练数据不足的问题.

1.3 视觉 Transformer

基于自注意力机制的架构, 尤其是 Transformer[30], 已成为自然语言处理领域的首选模型. 其中根据其能构建全局依赖性的特点开发的 BERT、Ro-BERT 和 GPTv1-3 模型取得良好的效果. 近年来, 受该架构在自然语言处理任务中成功应用的启发, 大量基于 Transformer 的模型逐渐被用来处理计算机视觉任务[31]. 在图像分类任务中, ViT[32]取消对 CNN 的依赖, 使用基于图像块序列的纯 Transformer 架构, 在 ImageNet 分类上表现最优. Chen 等人[33]提出一种基于预训练的 Transformer 模型, 用于解决图像超分辨率和去噪等不同的图像处理任务; 在医学图像分割任务中, UTNet[34]和 TransFuse 架构[35]均利用 CNN 和 Transformer 混合的结构实现精准的医学图像分割. 其中 UTNet[34]在编码器和解码器中应用自注意力模块, 以最小的开销捕捉不同尺度的长距离依赖关系, 实现端到端的心脏磁共振图像分割; 相似的, TransFuse 架构[35]也将 Transformer 和 CNN 以并行的方式结合在一起, 提高全局背景建模效率的同时又保持对低层次细节的学习, 从而在息肉、皮肤病变、髋关节和前列腺的分割数据集上取得较先进的实验结果; 针对深度伪造的人脸在互联网上广泛传播的情况, Khan 等人[36]提出一个具有增量学习功能的 Transformer 模型, 用于检测深度伪造的视频.

尽管近段时间以来 Transformer 已经在上述视觉领域中出现并取得一定的成果, 然而在图像篡改检测方法的应用上仍未被充分开发. 因此, 受其在图像分类与分割任务中成功应用的启发, 本文提出在编码器中引入 Transformer 架构以提取全局上下文依赖关系, 从而形成局部-全局特征增强模块, 并作为解码器的输入进一步细化网络的检测结果.

2 预训练驱动的多模态边界感知视觉 Transformer

本文提出的预训练驱动的多模态边界感知视觉 Transformer 主要由 4 个模块组成: (1) 频域模态 (frequency domain modality, FDM). (2) 基于预训练的局部-全局特征增强模块. (3) 边界感知模块 (boundary awareness module, BAM). (4) 渐进式语义生成模块 (progressive semantic generation module, PSGM). 总体网络结构如图 2 所示. 本文首先将多模态图像作为主干编码网络 CNN 的输入, 并通过预训练技术对其进行参数初始化, 在卷积层提取的高级语义特征后, 引入视觉 Transformer 模块进一步提取空间上下文的长远依赖关系. 其次, 将局部-全局特征增强模块输出的特征作为不同孔洞卷积操作的输入, 形成多尺度伪造特征. 然后, 将空间域卷积特征送入 Scharr 卷积层和边界残差块以生成篡改区域边界感知的特征图. 最后, PSGM 引导网络生成最终的检测结果图 **S**.

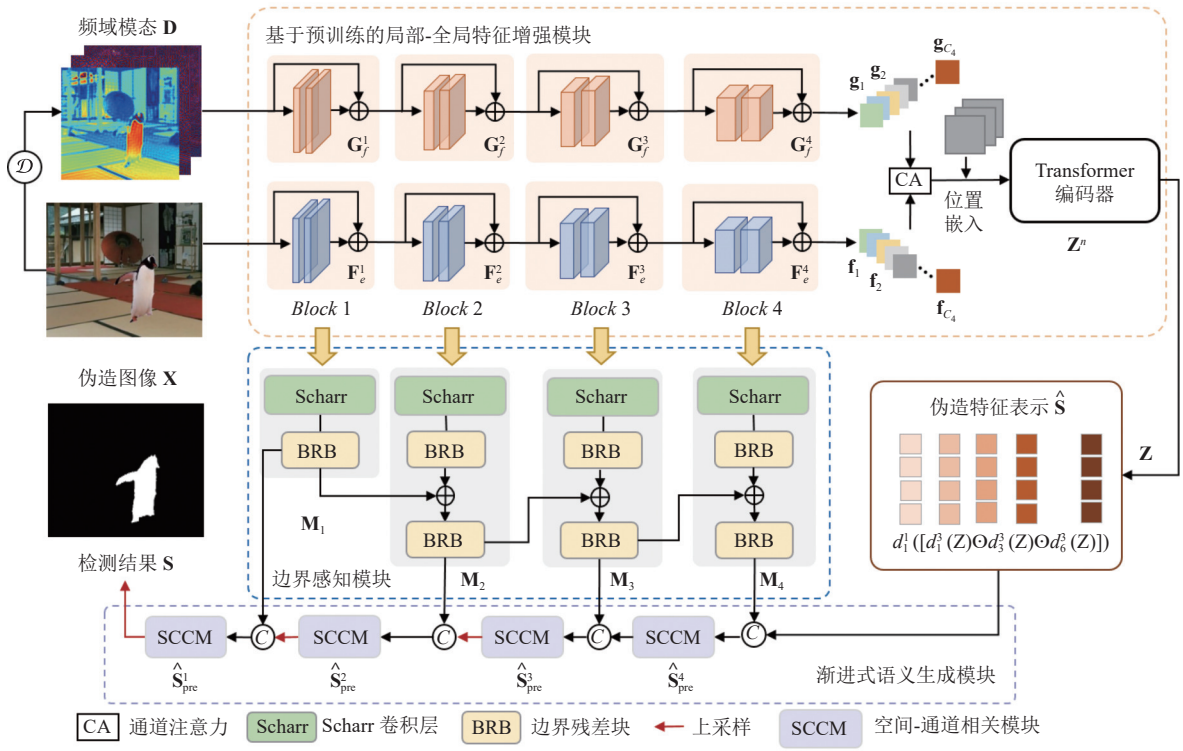


图 2 预训练驱动的多模态边界感知视觉 Transformer 网络结构图

2.1 频域模态

研究表明图像频域模态有利于感知篡改图像中的伪影信息, 即便是经过压缩操作 (例如 JPEG 压缩) 后不易发现的细微操纵线索[29,37]. 受人脸伪造检测中频域特征的启发[38,39], 我们引入频域模态作为图像 RGB 空间域信息的补充以挖掘伪造特征. 如图 3 所示, 将空间域图像 $\mathbf{X}$ 作为输入, 首先沿着空间维度应用离散余弦变换 $\mathcal{D}$ (discrete cosine transform, DCT) 将其从 RGB 域转换到频域并获得频谱表示 $\mathcal{D}(\mathbf{X})\in\mathbb{R}^{H\times W\times 1}$. 得益于 DCT 的特性, 低频响应位于频域分布 $\mathcal{D}(\mathbf{X})$ 的左上角, 而高频响应位于右下角. 然后, 我们手动设计 N 个二进制基础滤波器 $\{\mathbf{f}^i_{\text{base}}\}_{i=1}^N$, 从而明确地将频域划分为低、中、高频频段. 此外, 为自适应地选择感兴趣的频域信息并捕获伪造模式, 除基础滤波器外额外添加 3 个可学习的滤波器 $\{\mathbf{f}^i_{\text{learn}}\}_{i=1}^N$. 因此, 将频谱表示 $\mathcal{D}(\mathbf{X})$ 与组合滤波器相乘以模拟不同频带分量的依赖关系, 则图像频域模态可由公式 (1) 计算获得:

$$\mathbf{d}_i=\mathcal{D}^{-1}\{\mathcal{D}(\mathbf{X})\odot[\mathbf{f}^i_{\text{base}}+\sigma(\mathbf{f}^i_{\text{learn}})]\},\quad i=\{1,\ldots,N\} \tag{1}$$

其中, $\odot$ 为逐像素相乘 (element-wise product), $\sigma(x)=(1-\exp(-x))/(1+\exp(-x))$ 旨在压缩 x 在−1 和+1 之间的范围内, $\mathcal{D}^{-1}$ 表示逆 DCT. 本文的频带数 N=3, 低频带 $\mathbf{f}^1_{\text{base}}$ 为整个频谱的前 1/16, 中间频带 $\mathbf{f}^2_{\text{base}}$ 在频谱的 1/16 和 1/8 之间, 高频带 $\mathbf{f}^3_{\text{base}}$ 为频谱最后的 7/8. 最后, 沿着通道方向重新组合 $\{\mathbf{d}_i\}_{i=1}^3$ 以获得频域模态特征图 $\mathbf{D}\in\mathbb{R}^{H\times W\times 3}$.

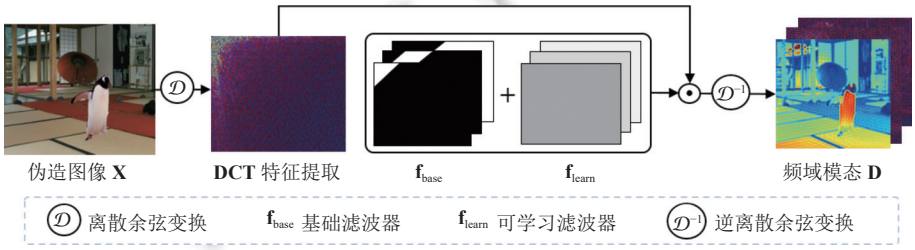


图 3 频域模态生成示意图

2.2 基于预训练的局部-全局特征增强模块

作为本文的主干编码网络, 该模块利用预训练技术, 融合经典的 CNN 编码器和 Transformer 模型, 并通过两个步骤达到特征增强的目的: (1) 基于预训练的多模态图像向量化和 (2) 局部-全局特征增强. 在图像向量化这一重要的步骤中, 图像被转换为一维序列的嵌入形式, 以适应后续序列转换操作. 局部-全局特征增强的设计是为了有效弥补图像向量化表示时忽略的全局语义信息, 并在图像块的级别上考虑块间的依赖性.

● 基于预训练的多模态图像向量化. 为降低因创建强注释带来的高成本, 缓解训练数据不足问题, 本文提出在特征增强模块中采用大规模 ImageNet 数据集进行模型预训练, 通过迁移训练好的参数, 代替随机化操作进行权重初始化, 以更好地学习输入图像的语义特征.

目前, 图像序列化方法主要包括: (1) CNN 编码器和 (2) 线性投影方法. 尽管线性投影方法在一些计算机视觉任务中取得一定成功, 但仍然存在一定缺陷, 即对图像数据量具有高度的依赖性. 为此, 我们选择通过 CNN 编码器进行图像序列化处理, 并将 ResNet50 网络作为局部-全局特征增强模块的主干网络. 其中, 给定一幅输入图像 $\mathbf{X}\in\mathbb{R}^{H\times W\times 3}$, 其高度为 H、宽度为 W、通道数为 C, 假设 $\{\mathbf{F}_e^i\in\mathbb{R}^{H_i\times W_i\times C_i}, i=1,\ldots,4\}$ 为 ResNet50 编码网络中第 i 个 ResNet 块 (*Block*) 输出的特征图, 其对应的特征图尺寸分别为 [H/4, W/4, 256]、[H/8, W/8, 512]、[H/16, W/16, 1024] 和 [H/16, W/16, 2048]. 同样, 图像频域模态也经过相同的主干网络. 因此, 针对图像 RGB 模态和频域模态, CNN 编码器的主干网络 ResNet50 最终产生的特征分别为 $\mathbf{F}_e^4=\{\mathbf{f}_i\}_{i=1}^{C_4}\in\mathbb{R}^{H_4\times W_4\times C_4}$ 和 $\mathbf{G}_f^4=\{\mathbf{g}_i\}_{i=1}^{C_4}\in\mathbb{R}^{H_4\times W_4\times C_4}$. 然后利用通道注意力 (channel attention, CA) 机制[40]对多模态特征进行融合, 同时为适应后续操作, 并利用 1×1 卷积层将特征图的个数由 4096 降低为 C_4. 最后将其展平 (flatten) 为一维 patch embedding, 并添加可学习的位置嵌入[41], 该嵌入被随机初始化以补偿被序列化破坏的空间信息, 从而生成最终的序列嵌入 $\mathbf{E}\in\mathbb{R}^{L\times C_4}, L=HW/256$, 如公式 (2):

$$\mathrm{E}=flatten(Att(cat(\mathbf{F}_e^4,\mathbf{G}_f^4)))\in\mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times C_4} \tag{2}$$

其中, *cat* 表示级联 (concatenation) 操作, $Att(\cdot)$ 表示通道注意力机制, $flatten(\cdot)$ 表示平铺函数.

● 局部-全局特征增强. 在图像序列化步骤中, 虽然基于 ResNet50 的 CNN 编码器输出的特征图包含丰富的局部空间信息和细节信息, 但仍缺少全局上下文信息. 考虑到伪造图像与自然图像存在的差异, 为更好地区分篡改区域与真实区域, 局部与全局等上下文信息对于目标区域的识别至关重要. 因此, 本文在主干网络 ResNet50 的最后一个 bottleneck 处引入 Transformer 编码器用于捕获整个伪造图像中的长远依赖关系, 以实现局部-全局特征增强的效果. 遵循现有的设计[30], Transformer 编码器是由 n 个堆叠编码器层组成 (n=4), 其每一层都由一个多头自注意力模块和一个多层感知器组成. 假设第 i 层的输入为 $\mathbf{Z}^{i-1}$ (特别的, $\mathbf{Z}^0\leftarrow\mathbf{E}$), 则其输出定义如下:

$$\mathbf{Z}^i=MSA(\mathbf{Z}^{i-1})\oplus MLP(MSA(\mathbf{Z}^{i-1})) \tag{3}$$

其中, $\oplus$ 表示逐像素相加 (element-wise addition) 操作, *MSA* 和 *MLP* 分别表示多头自注意力模块和多层感知器. 最终, 为进行下一阶段的伪造特征表示, 将最后一层 $\mathbf{Z}^n$ 的转换特征恢复为 2D 格式, 即 $\mathbf{Z}\in\mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times C_4}$.

为了在多个尺度上分割篡改区域, 本文将经过自注意力机制后的转换特征 $\mathbf{Z}$ 作为输入, 利用孔洞卷积操作以不同大小的感受野学习不同尺度伪造区域的上下文信息, 从而形成丰富的特征表示, 进一步提升网络的分割性能, 其可通过公式 (4) 计算得出:

$$\hat{\mathbf{S}}=d_1^1([cat(d_1^3(\mathbf{Z}),d_3^3(\mathrm{Z}),d_6^3(\mathbf{Z}))]) \tag{4}$$

其中, $d_r^s(\cdot)$ 表示扩张率为 r 且滤波器大小 $s\times s$ 的孔洞卷积函数, *cat* 为级联操作. 通过将具有不同感受野的增强特征图 $d_r^s(\mathbf{Z})$ 跨通道地拼接在一起, 生成伪造特征表示 $\hat{\mathbf{S}}$.

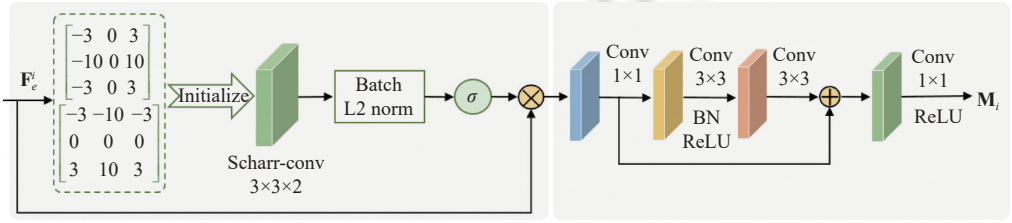
2.3 边界感知模块

尽管特征增强模块较好地提取图像浅层和深层的语义信息, 但因伪造图像中篡改区域的形状不规则、尺度变化多样以及边界模糊等因素, 给伪造图像的检测带来一定挑战. 为了缓解这一问题, 受文献 [24,26] 的启发, 本文专门设计 BAM 用于捕捉篡改区域周围的边界伪影. 然而, 如何为边界感知模块构建合适的输入是该模块设计的主要挑战. 代替传统简单的特征级联, 本文采用从浅到深的方式构建 BAM 中边界头 (Scharr 卷积层) 的输入, 从而达到浅层特征中细微边界伪影与深层特征充分提取与监督的目的.

尽管 Sobel 算子[24]可以有效地提取图像边缘, 但是对图像中较弱的边缘提取效果较差. 因此, 为充分地提取较弱的边缘, 增强与篡改边缘相关模式的感知, 本文提出一种基于 Scharr 算子的 Scharr 卷积层. 如图 4(a) 所示, 将来自不同 ResNet 块的特征作为 BAM 的输入, 第 i 个块的特征 $\mathbf{F}_e^i$ 首先通过 Scharr 卷积层, 并利用 Scharr 算子对其进行初始化. 此外, 为减少浅层特征中存在的误导信息, 本文随后引入边界残差块 (boundary residual block, BRB), 如图 4(b) 所示, 将该模块设计为残差结构并通过求和的方式与来自下一个块的对应部分进行特征组合. 为防止累积效应 (accumulation effect)[24], 在下一轮特征组合之前, 组合后的特征会再次经过一个 BRB (图 2 中的底部). 这种渐进式设计融合所有相邻的特征图, 有助于将边界细节信息从低层逐层传递到高层语义特征, 并在一定程度上抑制噪声信息. 因此, 在 BAM 中, 每一层级的最后一个 BRB 块将输出具有语义和边界信息的特征图 $\{\mathbf{M}_i \in \mathbb{R}^{H_i \times W_i \times 1}, i=1,\ldots,4\}$, 主要被用于生成篡改检测结果图. 具体操作定义如下:

$$\mathbf{M}_i = \begin{cases} \varphi(Scharr(\mathbf{F}_e^i)), & i=1 \\ \varphi(\mathbf{M}_{i-1} \oplus \varphi(Scharr(\mathbf{F}_e^i))), & i \in [2,3,4] \end{cases} \tag{5}$$

其中, $Scharr(\cdot)$ 和 $\varphi(\cdot)$ 分别表示 Scharr 卷积层和边界残差块.



(a) Scharr 卷积层 (b) 边界残差块 (BRB)

图 4 Scharr 卷积层和边界残差块的示意图

2.4 渐进式语义生成模块

受文献 [28] 的启发, 本文在 PSGM 中引入空间-通道相关模块 (spatial-channel correlation module, SCCM) 以在渐进式的路径中捕获空间和通道方面的相关性, 并赋予特征整体线索, 使网络能够集中应对篡改区域. 如图 2 所示, 本文通过模仿人类处理日常中复杂问题的方式, 利用具有完全监督的渐进机制, 避免直接以最精细尺度生成的伪造掩码. PSGM 联合上述的伪造特征表示 $\hat{\mathbf{S}}$ 和 BRB 输出的边界特征映射 $\mathbf{M}_i$ 作为输入, 通过 SCCM 对每个层级的明确监督, 实现由粗到细的掩码预测结构. SCCM 的预测结果可以通过公式 (6) 计算得出:

$$\hat{\mathbf{S}}_{\text{pre}}^{i-1} = \begin{cases} \xi(cat(\hat{\mathbf{S}}, \mathbf{M}_{i-1})), & i=5 \\ \xi(cat(\hat{\mathbf{S}}_{\text{pre}}^i, \mathbf{M}_{i-1})), & i=4 \\ \xi(cat(\tau(\hat{\mathbf{S}}_{\text{pre}}^i), \mathrm{M}_{i-i})), & i=2,3 \end{cases} \tag{6}$$

其中, ξ 表示 SCCM 模块, τ 是上采样操作 (例如双线性插值). 对于层级 1 (i=2) 和层级 2 (i=3), 当前层级上的特征 $\mathbf{M}_{i-1}$ 与前一个层级上采样 $\tau(\hat{\mathbf{S}}_{\text{pre}}^i)$ 相关联, 以产生一幅当前层级的掩码图 $\hat{\mathbf{S}}_{\text{pre}}^{i-1}$. 由于 $\mathbf{S}$ 是 PSGM 最后生成的检测结果图, 则与输入图像具有相同大小的特征图的最终预测结果可以直接表示为 $\mathbf{S}=\tau(\hat{\mathbf{S}}_{\text{pre}}^1)$.

图 5 展示 SCCM 的详细结构, 设 $\mathbf{I}$ 为输入特征, 利用函数 h 将输入 $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$ 转换为 $\mathbf{I}' \in \mathbb{R}^{HW/r^2 \times Cr^2}$, r 为下采样比例. 我们使用 1×1 卷积建立不同的函数 ρ, θ, ϕ 将 $\mathbf{I}'$ 转换为新的线性嵌入 $\mathbf{I}'_\rho=\rho(\mathbf{I}')$, $\mathbf{I}'_\theta=\theta(\mathbf{I}')$ 和 $\mathbf{I}'_\phi=\phi(\mathbf{I}')$. 通道注意力和空间注意力分别被表示为:

$$\begin{cases} \mathbf{Y}'_c = \mathbf{I}'_\rho \otimes \mathbf{A}_c = \mathbf{I}'_\rho \otimes Softmax(\mathbf{I}'^{\mathrm{T}}_\theta \mathbf{I}'_\phi) \\ \mathbf{Y}'_s = \mathbf{A}_s \otimes \mathbf{I}'_\rho = Softmax(\mathbf{I}'_\theta \mathbf{I}'^{\mathrm{T}}_\phi) \otimes \mathbf{I}'_\rho \end{cases} \tag{7}$$

其中, $\mathbf{A}_c$ 和 $\mathbf{A}_s$ 中的像素值分别表示通道图和空间图中相似性, $\otimes$ 为矩阵乘法 (matrix multiplication). 为生成和输入特征具有相同大小的特征表达 $\hat{\mathbf{I}}$, 分别通过函数 h 的逆变换, w_c 和 w_s 提高特征表达能力, 并引入两个可学习的参数 α_c 和 α_s. $\hat{\mathbf{I}}$ 的计算方式如下:

$$\hat{\mathbf{I}}=\mathbf{I}\oplus\alpha_c\cdot\omega_c(h^{-1}(\mathbf{Y}_c'))\oplus\alpha_s\cdot\omega_s(h^{-1}(\mathbf{Y}_s'))=\mathbf{I}\oplus\alpha_c\cdot\omega_c(\mathbf{Y}_c)\oplus\alpha_s\cdot\omega_s(\mathbf{Y}_s) \tag{8}$$

其中, $\oplus$ 为逐元素求和 (element-wise sum), 基于特征表示 $\hat{\mathbf{I}}$, 本文采用顺序为 *Conv-ReLU-Conv-Sigmoid* 的掩码生成块, 其中 *Conv* 是 3×3 卷积.

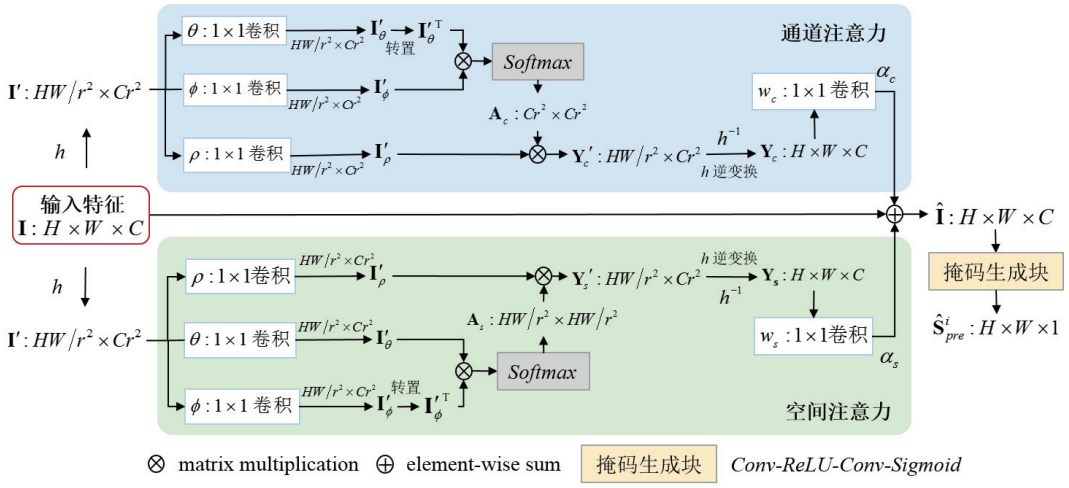


图 5 SCCM 示意图.

2.5 损失函数

由于伪造图像中被篡改的像素通常是少数的, 因此本文选择能够在极端不平衡数据中进行有效学习的 Dice 损失, 以最小化 **GT** (ground truth, GT) 和最终检测结果 **S** 之间的差异并计为 L_{seg}. 此外, PSGM 中逐步提供的特征图 $\hat{\mathbf{S}}_{\text{pre}}^i$ 能够引导网络最终生成语义丰富且更精准的检测结果, 从而提高网络的分割定位精度. 为此, 本文选择常用的交叉熵 (cross-entropy) 损失以减少 $\hat{\mathbf{S}}_{\text{pre}}^i$ 和其对应的真实掩码图 $\mathbf{S}_{\text{GT}}^i$ 之间的差距并计为 L_{pre}. 因此, 总的损失函数 L_{total} 可以定义为:

$$\begin{cases} L_{\text{total}}=L_{\text{seg}}+\dfrac{1}{4}\left(\sum_{i=1}^{4}L_{\text{pre}}^i\right) \\ L_{\text{seg}}=\phi_{\text{DICE}}(\mathbf{S},\mathbf{GT}),\ L_{\text{pre}}^i=\phi_{\text{CE}}(\hat{\mathbf{S}}_{\text{pre}}^i,\mathbf{S}_{\text{GT}}^i) \end{cases} \tag{9}$$

其中, $\hat{\mathbf{S}}_{\text{pre}}^i$ 表示 PSGM 中第 i 个层得到的预测结果. 此外, 通过对 **GT** 进行下采样得到 $\mathbf{S}_{\text{GT}}^i$, 以实现对每个预测掩码的全面监督, 其中 0 代表原始像素, 1 代表伪造像素. Dice 损失函数 ϕ_{DICE} 和交叉熵损失函数 ϕ_{CE} 分别定义如下:

$$\phi_{\text{DICE}}(\Theta_{\text{seg}})=1-\frac{2\cdot\sum_{j=1}^{W\times H}\mathbf{S}_j\cdot\mathbf{GT}_j}{\sum_{j=1}^{W\times H}\mathbf{S}_j^2+\sum_{j=1}^{W\times H}\mathbf{GT}_j^2} \tag{10}$$

$$\phi_{\text{CE}}(\Theta_{\text{pre}}^i)=-\sum_{j=1}^{N}\sum_{c\in\{0,1\}}\delta(\mathbf{S}_{\text{GT}}^i=c)\log p(\hat{\mathbf{S}}_{\text{pre}}^i=c) \tag{11}$$

其中, $\mathbf{GT}_j\in\{0,1\}$ 表示 **GT** 中第 j 个像素是否被篡改. 同理, $\mathbf{S}_j$ 表示 **S** 中第 j 个像素的概率值. N 为 $\hat{\mathbf{S}}_{\text{pre}}^i$ 的像素个数, δ 为指标函数, Θ_{seg}和Θ_{pre}^i 分别为最终分割结果图 **S** 和预测特征图 $\hat{\mathbf{S}}_{\text{pre}}^i$ 的参数集.

3 实验结果及分析

3.1 实验设置

3.1.1 数据集

我们使用 DEFACTO 数据集[42]对本文提出的模型进行预训练, 并在 4 个公开的图像篡改数据集-NIST16[43]、

COVERAGE[11]、Columbia[44]和 CASIA[45]上验证测试本文方法的有效性. 数据集的具体细节描述如下.

(1) 预训练数据集

● DEFACTO 是基于 MSCOCO[46]生成的合成数据集, 涵盖 3 种典型的伪造类型 (即拼接、复制-粘贴和移除). 本文从 DEFACTO 中选择 90 000 张篡改图像作为基础数据集, 用于预训练和消融实验方面的研究. 值得一提的是, 我们使用的基础数据集所包含的图像数量少于部分其他研究, 例如 SPAN (102 028 个样本)[21]、PSCC-Net (400 000 个样本)[28]和 SAT (98 779 个样本)[5]. 该数据集的训练-测试比率设置为 9:1.

(2) 标准测试集

● NIST16 是一个由 564 个样本组成的挑战性数据集, 涉及拼接、复制-粘贴和移除 3 种伪造类型. 此外, 所有图像都经过后处理, 使隐藏在数据中的篡改线索更难被发现. 且该数据集提供 ground truth.

● COVERAGE 专注于复制-粘贴篡改, 是一个包含 100 张伪造图像的小数据集. 所有图像都经过后处理以去除视觉伪造痕迹, 并提供 ground truth.

● Columbia 提供 180 幅带有边缘掩码的拼接图像, 其 ground truth 是由我们基于相应边缘掩码生成的.

● CASIA 主要关注拼接和复制-粘贴图像, 其所选择的篡改区域小而精细, 且部分伪造图像经过滤波和模糊等后处理操作. 它分为用于训练的 CASIA v2.0 (5 123 个样本) 和用于测试的 CASIA v1.0 (921 个样本) 两个版本. 两者都提供用于评估的二进制 ground truth.

为公平的比较, 我们遵循 RGB-N[18]中相同的训练-测试比率设置. 同时为避免任何数据泄露, 对于用于训练 (测试) 的伪造图像, 它们的源图像不包含在测试 (训练) 集中. 具体划分方式如表 1 所示.

表 1 预训练与标准数据集上图像训练-测试数量的划分

数据集	DEFACTO[42]	NIST16[43]	COVERAGE[11]	Columbia[44]	CASIA[45]
Training	80 000	404	75	—	5 123 (v2.0)
Testing	10 000	160	25	180	921 (v1.0)

3.1.2 网络实现细节

本文实验由开源的 PyTorch 深度学习框架实现并使用单个 NVIDIA GeForce RTX 3090 进行训练. 考虑到服务器的配置, 我们将图片大小调整为 512×512. 模型训练过程中, 采用初始学习率为 0.000 1 的 Adam 来优化网络模型. 当验证集的损失未能在 10 个 epoch 内下降时, 则学习率将下降 10%, 直至达到 1E–8. 本文算法的骨干网络在 ImageNet 上进行预训练, 其所有参数共经历 500 个 epoch 的微调.

3.1.3 评价指标

本文使用图像篡改检测任务中常用的两种评估指标来验证算法的性能, 主要包括: 像素级 F_1 分数和 AUC (area under the receiver operating characteristic curve). F_1 分数是用于图像篡改检测的像素级别评估指标且其值越大越好. F_1 分数和 AUC 的取值范围为 [0, 1]. 本文将篡改像素视为正样本, 真实像素视为负样本. 对于每个测试图像, 我们会改变不同的阈值, 并获得最大的 F_1 分数作为最终结果.

3.2 网络模型有效性消融实验

本节主要在 DEFACTO 数据集上进行消融实验研究, 以验证本文使用的预训练策略和各个模块组件的有效性. 我们采取建立不同模块组合形式来评估相应模块对网络模型的贡献和检测能力. 具体组合形式说明如下.

● Baseline. 该 Baseline 模型主要包含基于图像 RGB 模态的 ResNet50 主干编码网和伪造特征表示模块, 利用顺序为 *Conv-ReLU-Conv-Sigmoid* 的掩码生成块代替渐进式语义生成模块中的空间-通道相关模块. 此外并不包含预训练策略.

● Baseline+FDM. 该组合表示在 Baseline 模型的基础上, 引入图像频域模态, 与图像 RGB 模态共同作为主干编码网络的输入, 从多模态的角度更好地学习和捕捉图像的篡改伪影信息.

● Baseline+FDM+P. 该组合表示在 Baseline+FDM 模型的基础上, ResNet50 主干编码网络采用 ImageNet 数据

集上预训练的网络参数模型, 代替传统随机初始化的方法更好地学习图像特征信息.

● Baseline+FDM+P+Trans. 该组合表示在 Baseline+FDM+P 组合的基础上加入 Transformer 编码器, 因此与 ResNet50 编码网形成局部-全局特征增强模块, 从而进一步提取空间上下文的依赖关系.

● Baseline+FDM+P+Trans+BAM. 该组合表示在 Baseline+FDM+P+Trans 组合的基础上增加边界感知模块, 其目的是捕捉篡改区域周围细微的边界伪影以加强篡改区域边界的特征感知能力, 使网络更关注伪造与真实区域间不确定分类区域的学习.

● Baseline+FDM+P+Trans+BAM+PSGM. 该组合为本文提出的网络结构, 将基于预训练的局部-全局特征增强模块得到的伪造特征和边界感知模块输出的特征一并送入渐进式语义生成模块, 通过空间-通道相关性模块渐进式的计算空间和通道特征映射之间的相关性, 以增强感兴趣区域的表示, 对每个层级特征的进行明确监督, 实现由粗到细的掩码预测结构.

所有模型都使用相同的设置进行训练, 结果如表 2 所示. Baseline+FDM 通过引入图像频域模态在 AUC 和 F_1 分数上均有提高, 即验证了频域模态能够检测 RGB 域中不可见的伪造痕迹, 并作为互补信息与 RGB 特征结合有效提高模型的检测精度. 与 Baseline+FDM 相比, Baseline+FDM+P 因使用预训练的策略, 代替传统随机化操作进行权重初始化, 更好地帮助模型学习了输入图像的语义特征, 因此, AUC 和 F_1 分数分别提高 1.3% 和 1.6%. 基于此, Transformer 自注意力机制引入后的结果表明, 在多模态融合的基础上提取全局上下文对改善伪造区域的检测至关重要. 此外, Baseline+FDM+P+Trans+BAM 通过增加边界感知模块更多地关注图像噪声信息而不是语义图像内容, 从而增强网络模型对浅层特征中边界伪影与深层特征的学习能力, 提升网络模型在边界处的分割性能. 最后, 因本文在伪造特征表示和边界感知模块后加入渐进式语义生成模块, 使 Baseline+FDM+P+Trans+BAM+PSGM 组合的网络结构从空间和通道两个角度逐渐增强不同尺度特征图的相关性, 并逐步关注篡改区域取得最优的 AUC 和 F_1 分数. 因此, 消融实验结果进一步验证本文中各模块的有效性.

表 2 DEFACTO 数据集上各模块组合的消融实验对比结果

模型变体	模块					AUC	F_1
	FDM	Pre-trained	Transformer	BAM	PSGM		
Baseline						0.956	0.856
Baseline+FDM	√					0.965	0.871
Baseline+FDM+P	√	√				0.978	0.887
Baseline+FDM+P+Trans	√	√	√			0.987	0.909
Baseline+FDM+P+Trans+BAM	√	√	√	√		0.992	0.932
Baseline+FDM+P+Trans+BAM+PSGM	√	√	√	√	√	**0.996**	**0.940**

3.3 与其他方法对比的定量实验结果

为证明本文提出的预训练驱动的多模态边界感知视觉 Transformer 在图像篡改检测方面的优势, 我们在 4 个基准图像库 (NIST16[43]、COVERAGE[11]、Columbia[44]和 CASIA[45]) 上将本文方法与 3 种经典的无监督方法 ELA[47]、NOI1[48]和 CFA1[49], 以及最新的深度网络模型 ManTra-Net[17]、J-LSTM[15]、H-LSTM[16]、RGB-N[18]、GSR-Net[26]、SPAN[21]、PSCC-Net[28]和 SAT[5]进行比较. 为保证客观和公平的比较, 我们采用两种不同的实验设置: (1) Pre-trained. 该预训练设置表示在 DEFACTO 数据集上进行训练, 并在完整的测试数据集上进行评估. (2) Fine-tuned. 该微调设置利用测试数据集的训练部分进一步微调预训练模型, 并在其测试部分进行评估.

● Pre-trained. 表 3 显示在 Pre-trained 实验设置下, 不同方法在 4 个标准数据集上定量的对比结果. 从表 3 中可以看出, 本文方法在 NIST16、CASIA 和 Columbia 上实现最佳 AUC 检测结果, 在 COVERAGE 上排名第 2. 其中, 与 PSCC-Net 相比, 本文方法在 NIST16 (↑3.7%)、Columbia (↑0.9%) 和 CASIA (↑1.4%) 数据集上的 AUC 均有提高. 这表明, 我们方法中的频域模态具有捕获篡改特征的优越能力, 并可以很好地推广到不同质量的图像篡改数据集. 同时也验证了预训练驱动下本文多模态特征提取的有效性. 尽管在 COVERAGE 上我们比 PSCC-Net 获得

1.9% 的收益, 但是未能在 Coverage 上取得最佳性能, 其原因可能是我们的训练数据相对不够完善. 然而, 本文方法在 4 个数据集上的 AUC 平均值排名第 1, 即验证了与其他网络相比具有较好的泛化能力.

表 3 Pre-trained 设置下 AUC 的定量比较

方法	NIST16[43]	COVERAGE[11]	Columbia[44]	CASIA[45]	Mean
ManTra-Net[17]	0.795	0.819	0.824	0.817	0.814
SPAN[21]	0.840	**0.922**	0.936	0.797	0.874
PSCC-Net[28]	0.855	0.847	0.982	0.829	0.878
本文方法	**0.892**	0.866	**0.991**	**0.843**	**0.898**

● Fine-tuned. 我们进一步利用标准测试数据集中的训练数据对预训练模型进行微调, 通过不同训练数据集的交叉验证在每个测试数据集上选择最佳的微调模型. 表 4 给出在 Fine-tuned 实验设置下本文方法与其他对比方法的像素级 AUC 和 F_1 分数的定量比较, 其中, "—"表示比较方法的原论文没有提供该项实验结果. 与无监督方法相比, 本文方法显著提升篡改检测性能, 表明基于深度学习的有监督架构具有一定优势. 与一些最新的深度网络方法相比, 本文方法在 NIST16、COVERAGE 和 CASIA 数据集上均取得最佳性能. 一种可能的解释是, 这些方法尽管利用一些微妙的网络模块来提取伪造区域的空间信息, 但忽略了图像的全局和边界信息, 导致在具有丰富语义信息的 NIST16、COVERAGE 和 CASIA 数据集图像上泛化性不佳. 同时, 尽管本文方法在 NIST16 的 AUC 与 PSCC-Net 相同, 但相比于 PSCC-Net (400 000 个样本) 和 SAT (98 779 个样本) 方法的预训练样本数量, 本文能够在相对较小规模 (90 000 个样本) 的训练数据上获得更好的微调结果. 尽管数据规模较小, 但本文提出的多模态嵌入形式的输入及视觉 Transformer 模块对精度的提升具有主要贡献作用. 其中, 与 SAT 相比, 本文方法在 NIST16、COVERAGE 和 CASIA 上的 AUC 分别提高 0.6%、0.3% 和 4.3%, F_1 分数分别提高 3.6%、1.1% 和 3.1%. 此外, 本文方法在 3 个数据集上相应的 AUC 和 F_1 分数平均值排名第一. 具体来说, 与次优模型相比, 我们的模型在 AUC 和 F_1 分数方面平均超出 1.8% 和 2.6%, 即验证所提出的模块和预训练策略的有效性.

表 4 Fine-tuned 设置下与其他方法比较的定量结果

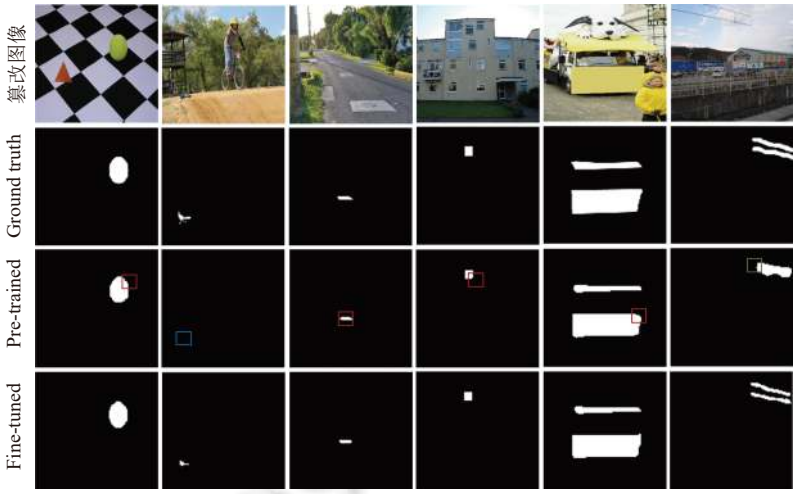
方法	训练设置	NIST16[43]		COVERAGE[11]		CASIA[45]		Mean	
		AUC	F_1	AUC	F_1	AUC	F_1	AUC	F_1
ELA[47]	unsupervised	0.429	0.236	0.583	0.222	0.613	0.214	0.542	0.224
NOI1[48]	unsupervised	0.487	0.285	0.587	0.269	0.612	0.263	0.562	0.272
CFA1[49]	unsupervised	0.501	0.174	0.485	0.190	0.522	0.207	0.503	0.190
J-LSTM[15]	fine-tuned	0.764	—	0.614	—	—	—	0.689	—
H-LSTM[16]	fine-tuned	0.794	—	0.712	—	—	—	0.753	—
RGB-N[18]	fine-tuned	0.937	0.722	0.817	0.437	0.795	0.408	0.850	0.522
GSR-Net[26]	fine-tuned	0.945	0.736	0.768	0.489	0.796	0.574	0.836	0.600
SPAN[21]	fine-tuned	0.961	0.582	0.937	0.558	0.838	0.382	0.912	0.507
PSCC-Net[28]	fine-tuned	0.996	0.819	0.941	0.723	0.875	0.554	0.937	0.699
SAT[5]	fine-tuned	0.990	0.878	0.985	0.843	0.843	0.592	0.939	0.771
本文方法	fine-tuned	**0.996**	**0.914**	**0.988**	**0.854**	**0.886**	**0.623**	**0.957**	**0.797**

3.4 定性可视化结果

(1) NIST16 数据集上有无微调操作的定性实验结果

图 6 展示在 NIST16 数据集上本文提出的模型在 Pre-trained 和 Fine-tuned 实验设置下的定性实验结果. 与预训练模型相比, Fine-tuned 实验设置下本文方法可以从训练有素的模型中获得更精细的分割掩码. 微调模型的优势可以分为 3 种情况: ① 伪造区域的边界处预测, 如图中红色的矩形框. ② 小目标区域的检测, 如图中第 2 列的蓝色

矩形框标记的篡改区域 (动物鸟), Pre-trained 模型下并没有识别出篡改区域, 而 Fine-tuned 模型能够定位出动物鸟的主要身体区域. ③ 相似伪造区域的检测, 如图中绿色矩形框标记的篡改区域, 无微调操作的 Pre-trained 模型将两个伪造区域识别为一个整体, 导致出现一定程度的误定位现象. 因此, 利用预训练方法对模型进行微调, 能够使模型预测出更精准的伪造掩码, 同时保持对篡改区域边界处和小目标区域伪造的检测精度.



(a) 拼接篡改 (b) 复制-粘贴篡改 (c) 移除篡改

图 6 NIST16 数据集上的定性可视化结果

(2) NIST16 数据集上不同模块组合的定性实验结果

图 7 展示本文方法与不同模块组合的检测结果. 从上到下的结果分别为 NIST16 数据集中拼接、复制-粘贴和移除篡改图像. 由检测结果可知, 在 Baseline 模型的基础上, 利用多模态输入和预训练策略基本能够完成伪造区域的整体定位, 但边界处检测效果较差一些 (如图中第 3 列标记的红色矩形). 而在 Baseline+FDM+P 引入 Transformer 模块和边界感知 BAM 模块后, 网络模型能够对伪造区域边界处的一些欠分割和误分割部分进行补充和微调, 其可能原因是所提出的边界感知模块能够通过 Scharr 层过滤冗余的空间响应, 学习到更稳定的边界线索 (如图中第 4 列和第 5 列拼接图像中人物头部和手部篡改区域, 复制-粘贴图像中下半部分的篡改区域). 此外, 在 Baseline+FDM+P+Trans+BAM 的基础上引入 PSGM 后, 即本文提出的方法最终通过利用边界感知特征图和局部-空间特征图中的空间和通道之间的相关性, 以渐进式的方式对伪造区域进行有效分割, 进一步提高检测精准度. 以上可视化结果证明多模态输入、Transformer 模块、BAM 模块和 PSGM 模块可以帮助分割网络构建对伪造区域全局、边界以及不同空间-通道间的上下文的理解, 提高分割网络对伪造区域的辨识能力.

(3) COVERAGE、CASIA 和 Columbia 数据集上本文方法的定性实验结果

图 8 展示本文方法的可视化检测分割结果. 从上到下, 每两列图像分别来源于 COVERAGE、CASIA 和 Columbia 数据集. 前两个图像库的结果均是微调后的模型结果, 然而 Columbia 图像库并没有进行微调, 因此, 其显示的为预训练实验设置下的可视化结果. 从视觉分析上, 我们可以观察到本文方法在不同类型的篡改手段的伪造图像检测中取得很好的分割效果. 定性实验结果表明该方法不仅能更准确地定位篡改区域, 而且可以形成更清晰的边界, 这得益于本文方法长远依赖关系的建模能力和边界敏感性. 例如, 在第 1 行和第 2 行复制-粘贴篡改的相似性物体分割中, 本文方法能够有效抑制真实区域的内容, 避免误定位现象; 在第 3 行和第 4 行不同尺度、伪造区域与周围背景对比度低的情况中, 本文方法能够有效排除背景真实区域的相似性干扰, 从而一定程度地避免欠分割现象, 并在边界处取得较好的检测结果; 在第 5 行和第 6 行伪造区域包含多种实例对象的不规则区域时, 本文方法能够精准地定位伪造区域, 从而取得较优的检测精度. 我们的模型同时利用了空间域和频域特征, 并捕获了边界信息和噪声分布, 因此可以更好地从一种数据集推广到另一种数据集.

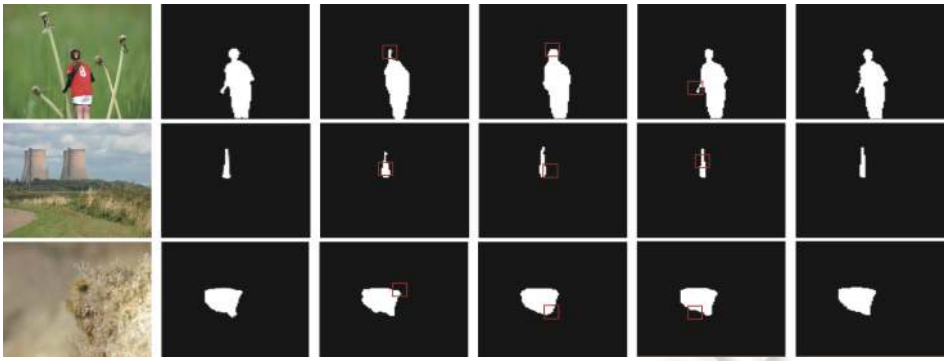


图 7 NIST16 数据集上不同模块组合的定性可视化结果

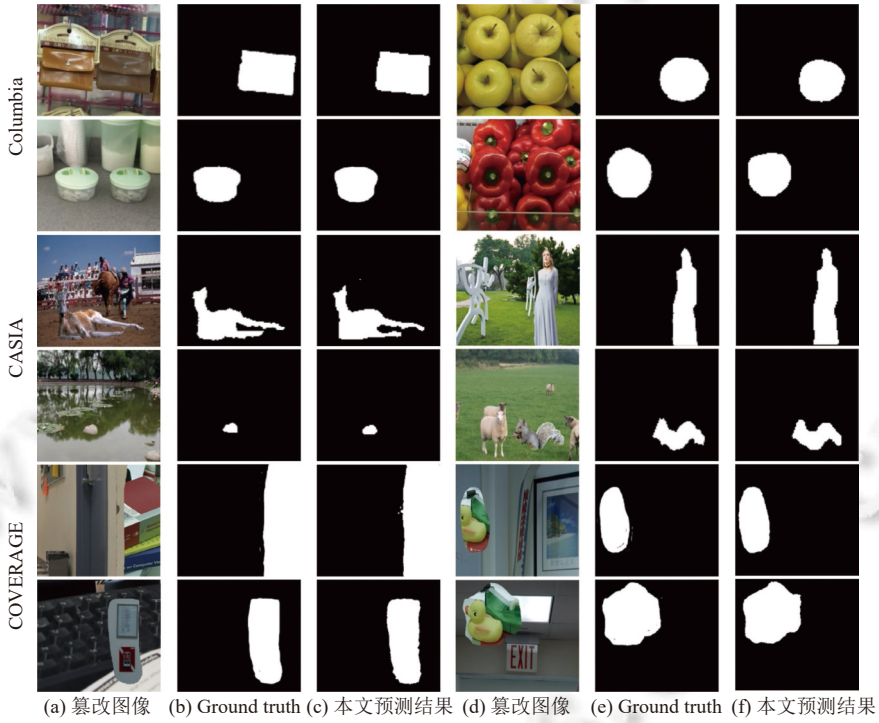


图 8 COVERAGE、CASIA 和 Columbia 数据集上的定性可视化结果

3.5 鲁棒性实验

为评估本文方法在检测任务方面的鲁棒性, 我们按照 SPAN[21]中的后处理失真操作设置, 利用 OpenCV 的内置函数在 NIST16 数据集上进行以下图像内容保留的后处理研究: (1) Resize: 将图像大小调整到不同的比例. (2) GaussianBlur: 应用 kernel 大小为 k 的高斯模糊. (3) GaussianNoise-添加标准偏差为 σ 的高斯噪声以及 (4) JPEGCompress: 执行质量因子为 q 的 JPEG 压缩. 图 9 显示在 Pre-trained 实验设置下, 本文方法与 ManTra-Net、SPAN 以及 PSCC-Net 的像素级 AUC 下的鲁棒性分析. 值得注意的是, 本文方法受 Resize 和 JPEGCompress 两种失真操作的影响相对较小, 表现出更强的鲁棒性能, 而对 GaussianNoise 失真操作表现得比较敏感. 其原因是本文将 DCT 应用于 RGB 空间域图像上, 通过收集一些包含被篡改区域的边界和细节的频域感知线索, 挖掘到细微的伪造伪影和压缩误差, 并通过 Transformer 模块提取了不同域之间的空间依赖关系. 因此, 本文方法可以为 Resize

和 JPEGCompress 两种失真操作提供更多的检测证据. 本文方法在所有的失真操作攻击上的表现一直优于 ManTra-Net[17]、SPAN[21]和 PSCC-Net[28], 因此表明其具有一定的鲁棒性.

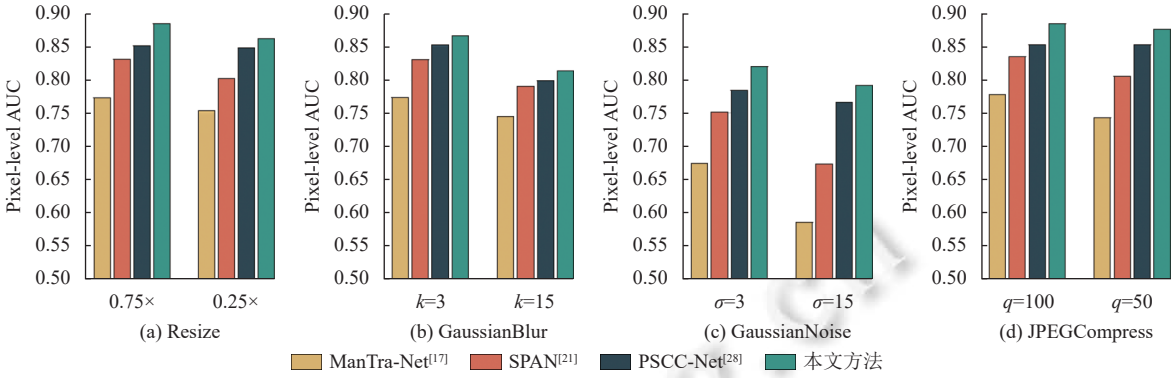


图 9　4 种不同失真操作的鲁棒性比较

4 结　论

面向图像篡改检测任务, 本文提出一种预训练驱动的多模态边界感知视觉 Transformer. 除原始图像 RGB 空间域以外, 首先引入基于离散余弦变换的图像频域信息共同作为模型主干网络的多模态输入. 其次, 利用图像分类领域中大规模的数据集 ImageNet 对局部-全局特征增强模块进行预训练, 从而缓解训练数据不足问题, 并在 CNN 编码器的 bottleneck 处引入 Transformer 模块, 从而提取全局上下文信息, 增强模型的表征能力. 然后边界感知模块通过 Scharr 卷积层和残差模块捕捉篡改区域周围的边界伪影, 以提升网络的边界分割性能. 最后本文将边界感知模块生成的特征图与伪造特征图作为监督信息送入渐进式语义生成模块, 利用 SSCM 模块更好地探索空间和通道方面的相关性, 以渐进的方式逐级生成最终的检测结果图. 实验结果表明, 本文提出的方法在不同基准数据集上的检测性能均优于目前先进的方法.

随着深度学习技术的持续发展, 真实场景中伪造手段必然越来越复杂且呈多样性, 给图像篡改检测带来更多新挑战. 同时人脸深度伪造 (DeepFake) 检测也是目前众多学者的研究方向之一. 展望未来, 为遏制图像篡改行为, 我们将继续在基准图像库的构建、模型泛化能力及鲁棒性的提升、深度伪造检测方面进行探索研究.

References:

[1] Li XR, Ji SL, Wu CM, Liu ZG, Deng SG, Cheng P, Yang M, Kong XW. Survey on deepfakes and detection techniques. Ruan Jian Xue Bao/Journal of Software, 2021, 32(2): 496–518 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6140.htm [doi: 10.13328/j.cnki.jos.006140]

[2] Li XL, Yu NH, Zhang XP, Zhang WM, Li B, Lu W, Wang W, Liu XL. Overview of digital media forensics technology. Journal of Image and Graphics, 2021, 26(6): 1216–1226 (in Chinese with English abstract). [doi: 10.11834/jig.210081]

[3] Verdoliva L. Media forensics and DeepFakes: An overview. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(5): 910–932. [doi: 10.1109/JSTSP.2020.3002101]

[4] Bik EM, Casadevall A, Fang FC. The prevalence of inappropriate image duplication in biomedical research publications. mBio, 2016, 7(3): e00809–16. [doi: 10.1128/mBio.00809-16]

[5] Zhuo L, Tan SQ, Li B, Huang JW. Self-adversarial training incorporating forgery attention for image forgery localization. IEEE Trans. on Information Forensics and Security, 2022, 17: 819–834. [doi: 10.1109/TIFS.2022.3152362]

[6] Huh M, Liu A, Owens A, Efros AA. Fighting fake news: Image splice detection via learned self-consistency. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 106–124. [doi: 10.1007/978-3-030-01252-6_7]

[7] Liu YQ, Zhu XB, Zhao XF, Cao Y. Adversarial learning for constrained image splicing detection and localization based on atrous convolution. IEEE Trans. on Information Forensics and Security, 2019, 14(10): 2551–2566. [doi: 10.1109/TIFS.2019.2902826]

[8] Kniaz VV, Knyaz VA, Remondino F. The point where reality meets fantasy: Mixed adversarial generators for image splice detection. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 20. [doi: 10.5555/3454287.3454307]

[9] Wu Y, Abd-Almageed W, Natarajan P. Deep matching and validation network: An end-to-end solution to constrained image splicing localization and detection. In: Proc. of the 25th ACM Int'l Conf. on Multimedia. Mountain View: ACM, 2017. 1480–1502. [doi: 10.1145/3123266.3123411]

[10] Wu Y, Abd-Almageed W, Natarajan P. BusterNet: Detecting copy-move image forgery with source/target localization. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 170–186. [doi: 10.1007/978-3-030-01231-1_11]

[11] Wen BH, Zhu Y, Subramanian R, Ng TT, Shen XJ, Winkler S. COVERAGE—A novel database for copy-move forgery detection. In: Proc. of the 2016 IEEE Int'l Conf. on Image Processing. Phoenix: IEEE, 2016. 161–165. [doi: 10.1109/ICIP.2016.7532339]

[12] Islam A, Long CJ, Basharat A, Hoogs A. DOA-GAN: Dual-order attentive generative adversarial network for image copy-move forgery detection and localization. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 4675–4684. [doi: 10.1109/CVPR42600.2020.00473]

[13] Zhu XS, Qian YJ, Zhao XF, Sun B, Sun Y. A deep learning approach to patch-based image inpainting forensics. Signal Processing: Image Communication, 2018, 67: 90–99. [doi: 10.1016/j.image.2018.05.015]

[14] Salloum R, Ren YZ, Kuo CCJ. Image splicing localization using a multi-task fully convolutional network (MFCN). Journal of Visual Communication and Image Representation, 2018, 51: 201–209. [doi: 10.1016/j.jvcir.2018.01.010]

[15] Bappy JH, Roy-Chowdhury AK, Bunk J, Nataraj L, Manjunath BS. Exploiting spatial structure for localizing manipulated image regions. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 4980–4989. [doi: 10.1109/ICCV.2017.532]

[16] Bappy JH, Simons C, Nataraj L, Manjunath BS, Roy-Chowdhury AK. Hybrid LSTM and encoder-decoder architecture for detection of image forgeries. IEEE Trans. on Image Processing, 2019, 28(7): 3286–3300. [doi: 10.1109/TIP.2019.2895466]

[17] Wu Y, AbdAlmageed W, Natarajan P. ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 9535–9544. [doi: 10.1109/CVPR.2019.00977]

[18] Zhou P, Han XT, Morariu VI, Davis LS. Learning rich features for image manipulation detection. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1053–1061. [doi: 10.1109/CVPR.2018.00116]

[19] Fridrich J, Kodovsky J. Rich models for steganalysis of digital images. IEEE Trans. on Information Forensics and Security, 2012, 7(3): 868–882. [doi: 10.1109/TIFS.2012.2190402]

[20] Zhu Y, Chen CF, Yan G, Guo YC, Dong YF. AR-Net: Adaptive attention and residual refinement network for copy-move forgery detection. IEEE Trans. on Industrial Informatics, 2020, 16(10): 6714–6723. [doi: 10.1109/TII.2020.2982705]

[21] Hu XF, Zhang ZH, Jiang ZY, Chaudhuri S, Yang ZH, Nevatia R. SPAN: Spatial pyramid attention network for image manipulation localization. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 312–328. [doi: 10.1007/978-3-030-58589-1_19]

[22] Li HD, Huang JW. Localization of deep inpainting using high-pass fully convolutional network. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 8300–8309. [doi: 10.1109/ICCV.2019.00839]

[23] Yang C, Li HZ, Lin FT, Jiang B, Zhao H. Constrained R-CNN: A general image manipulation detection model. In: Proc. of the 2020 IEEE Int'l Conf. on Multimedia and Expo. London: IEEE, 2020. 1–6. [doi: 10.1109/ICME46284.2020.9102825]

[24] Chen XR, Dong CB, Ji JQ, Cao J, Li XR. Image manipulation detection by multi-view multi-scale supervision. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 14165–14173. [doi: 10.1109/ICCV48922.2021.01392]

[25] Wang XY, Wang H, Niu SZ, Zhang JW. Detection and localization of image forgeries using improved mask regional convolutional neural network. Mathematical Biosciences and Engineering, 2019, 16(5): 4581–4593. [doi: 10.3934/mbe.2019229]

[26] Zhou P, Chen BC, Han XT, Najibi M, Shrivastava A, Lim SN, Davis L. Generate, segment, and refine: Towards generic manipulation segmentation. Proc. of the 2020 AAAI Conf. on Artificial Intelligence, 2020, 34(7): 13058–13065. [doi: 10.1609/aaai.v34i07.7007]

[27] Chen S, Yao TP, Chen Y, Ding SH, Li JL, Ji RR. Local relation learning for face forgery detection. Proc. of the 2021 AAAI Conf. on Artificial Intelligence, 2021, 35(2): 1081–1088. [doi: 10.1609/aaai.v35i2.16193]

[28] Liu XH, Liu YJ, Chen J, Liu XM. PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. IEEE Trans. on Circuits and Systems for Video Technology, 2022, 32(11): 7505–7517. [doi: 10.1109/TCSVT.2022.3189545]

[29] Luo YC, Zhang Y, Yan JC, Liu W. Generalizing face forgery detection with high-frequency features. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 16313–16321. [doi: 10.1109/CVPR46437.2021.01605]

[30] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010. [doi: 10.5555/3295222.3295349]

[31] Li YQ, Li CZ, Liu RQ, Si WX, Jin YM, Heng PA. Semi-supervised spatiotemporal Transformer networks for semantic segmentation of surgical instrument. Ruan Jian Xue Bao/Journal of Software, 2022, 33(4): 1501–1515 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6469.htm [doi: 10.13328/j.cnki.jos.006469]

[32] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021. 1–21.

[33] Chen HT, Wang YH, Guo TY, Xu C, Deng YP, Liu ZH, Ma SW, Xu CJ, Xu C, Gao W. Pre-trained image processing transformer. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 12294–12305. [doi: 10.1109/CVPR46437.2021.01212]

[34] Gao YH, Zhou M, Metaxas DN. UTNet: A hybrid transformer architecture for medical image segmentation. In: Proc. of the 24th Int'l Conf. on Medical Image Computing and Computer-assisted Intervention. Strasbourg: Springer, 2021. 61–71. [doi: 10.1007/978-3-030-87199-4_6]

[35] Zhang YD, Liu HY, Hu Q. Transfuse: Fusing transformers and CNNs for medical image segmentation. In: Proc. of the 24th Int'l Conf. on Medical Image Computing and Computer-assisted Intervention. Strasbourg: Springer, 2021. 14–24. [doi: 10.1007/978-3-030-87193-2_2]

[36] Khan SA, Dai H. Video transformer for deepfake detection with incremental learning. In: Proc. of the 29th ACM Int'l Conf. on Multimedia. ACM, 2021. 1821–1828. [doi: 10.1145/3474085.3475332]

[37] Wang JK, Wu ZX, Ouyang WH, Han XT, Chen JJ, Jiang YG, Li SN. M2TR: Multi-modal multi-scale transformers for DeepFake detection. In: Proc. of the 2022 Int'l Conf. on Multimedia Retrieval. Newark: ACM, 2022. 615–623. [doi: 10.1145/3512527.3531415]

[38] Qian YY, Yin GJ, Sheng L, Chen ZX, Shao J. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 86–103. [doi: 10.1007/978-3-030-58610-2_6]

[39] Wang JK, Wu ZX, Chen JJ, Han XT, Shrivastava A, Lim SN, Jiang YG. ObjectFormer for image manipulation detection and localization. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 2354–2363. [doi: 10.1109/CVPR52688.2022.00240]

[40] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141. [doi: 10.1109/CVPR.2018.00745]

[41] Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN. Convolutional sequence to sequence learning. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: JMLR.org, 2017. 1243–1252. [doi: 10.5555/3305381.3305510]

[42] Mahfoudi G, Tajini B, Retraint F, Morain-Nicolier F, Dugelay JL, Pic M. DEFACTO: Image and face manipulation dataset. In: Proc. of the 27th European Signal Processing Conf. A Coruna: IEEE, 2019. 1–5. [doi: 10.23919/EUSIPCO.2019.8903181]

[43] Guan HY, Kozak M, Robertson E, Lee YY, Yates AN, Delgado A, Zhou DL, Kheyrkhah T, Smith J, Fiscus J. MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In: Proc. of the 2019 IEEE Winter Applications of Computer Vision Workshops. Waikoloa: IEEE, 2019. 63–72. [doi: 10.1109/WACVW.2019.00018]

[44] Hsu YF, Chang SF. Detecting image splicing using geometry invariants and camera characteristics consistency. In: Proc. of the 2006 IEEE Int'l Conf. on Multimedia and Expo. Toronto: IEEE, 2006. 549–552. [doi: 10.1109/ICME.2006.262447]

[45] Dong J, Wang W, Tan TN. CASIA image tampering detection evaluation database. In: Proc. of the 2013 IEEE China Summit and Int'l Conf. on Signal and Information Processing. Beijing: IEEE, 2013. 422–426. [doi: 10.1109/ChinaSIP.2013.6625374]

[46] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common objects in context. In: Proc. of the 13th European Conf. on Computer Vision. Zurich: Springer, 2014. 740–755. [doi: 10.1007/978-3-319-10602-1_48]

[47] Krawetz N, Solutions HF. A picture's worth. Hacker Factor Solutions, 2007, 6(2): 1–31.

[48] Mahdian B, Saic S. Using noise inconsistencies for blind image forensics. Image and Vision Computing, 2009, 27(10): 1497–1503. [doi: 10.1016/j.imavis.2009.02.001]

[49] Ferrara P, Bianchi T, de Rosa A, Piva A. Image forgery localization via fine-grained analysis of CFA artifacts. IEEE Trans. on Information Forensics and Security, 2012, 7(5): 1566–1577. [doi: 10.1109/TIFS.2012.2202227]

附中文参考文献:

[1] 李旭嵘, 纪守领, 吴春明, 刘振广, 邓水光, 程鹏, 杨珉, 孔祥维. 深度伪造与检测技术综述. 软件学报, 2021, 32(2): 496–518. http://

www.jos.org.cn/1000-9825/6140.htm [doi: 10.13328/j.cnki.jos.006140]

[2] 李晓龙, 俞能海, 张新鹏, 张卫明, 李斌, 卢伟, 王伟, 刘晓龙. 数字媒体取证技术综述. 中国图象图形学报, 2021, 26(6): 1216–1226. [doi: 10.11834/jig.210081]

[31] 李耀仟, 李才子, 刘瑞强, 司伟鑫, 金玥明, 王平安. 面向手术器械语义分割的半监督时空Transformer网络. 软件学报, 2022, 33(4): 1501–1515. http://www.jos.org.cn/1000-9825/6469.htm [doi: 10.13328/j.cnki.jos.006469]

石泽男(1993－), 女, 博士, CCF 学生会员, 主要研究领域为计算机视觉, 医学图像分割, 多媒体取证.

张冬(1989－), 男, 博士, 主要研究领域为目标检测, 语义分割, 视频对象分割, 跨场景分割.

陈海鹏(1978－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为机器学习与视觉推理.

申铉京(1958－), 男, 博士, 教授, 博士生导师, 主要研究领域为医学图像分割, 多媒体取证, 光电及混合系统, 智能测量系统, 视频理解技术.