

# 基于 Convnext-Upernet 的图像篡改检测定位模型

胡林辉<sup>1),2),3)</sup> 陈保营<sup>1),2),3)</sup> 谭舜泉<sup>1),2),3)</sup> 李斌<sup>1),2)</sup>

<sup>1)</sup>(深圳市媒体信息安全重点实验室 广东 深圳 518060)

<sup>2)</sup>(广东省智能信息重点实验室 广东 深圳 518060)

<sup>3)</sup>(深圳大学计算机与软件学院 广东 深圳 518060)

**摘要** 在当前数字时代,假新闻、网络勒索等网络犯罪行为愈发猖獗,导致篡改图像产生的负面影响日益凸显. 鉴于此,检测与定位篡改图像已成为图像取证领域的关键任务. 近年来,深度学习技术在计算机视觉领域取得了显著进展,众多篡改检测模型亦逐渐应用该技术. 然而,现有模型大多需要在大量数据上进行预训练,且其鲁棒性和泛化能力相对较弱. 为解决上述问题,本研究采用在计算机视觉领域表现优异的纯卷积神经网络模型 Convnext 作为主干网络,并借助统一感知解析网络 Upernet 提取图像中的多尺度特征,构建了一种基于 Convnext-Upernet 的篡改检测定位模型. 在此基础上,本研究进一步运用自监督数据增强方法放大图像中的篡改痕迹,并利用与篡改检测定位任务无关的图像分类损失函数提高篡改图像检测定位的准确性. 本研究在当前主流的篡改检测定位数据集上进行了大规模实验证明,所提出的模型具有高效且精确的篡改检测定位能力. 相较于现有跨库性能最佳的 MVSS-Net++ 模型,本研究所提出的模型在检测定位性能上提高了 14.4%,泛化能力得到全面提升,并对常见的后处理操作展示出了强大的鲁棒性.

**关键词** 图像篡改检测定位;深度学习;卷积神经网络;统一感知解析网络;自监督数据增强  
**中图分类号** TP391 **DOI号** 10.11897/SP.J.1016.2023.02225

## Convnext-Upernet Based Deep-Learning Model for Image Forgery Detection and Localization

HU Lin-Hui<sup>1),2),3)</sup> CHEN Bao-Ying<sup>1),2),3)</sup> TAN Shun-Quan<sup>1),2),3)</sup> LI Bin<sup>1),2)</sup>

<sup>1)</sup>(Shenzhen Key Laboratory of Media Security, Shenzhen, Guangdong 518060)

<sup>2)</sup>(Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen, Guangdong 518060)

<sup>3)</sup>(College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, Guangdong 518060)

**Abstract** In the contemporary digital era, the proliferation of cybercrimes, such as fake news and online extortion, has led to increasingly prominent negative impacts caused by manipulated images. Consequently, detecting and locating manipulated images has become a critical task in the field of image forensics. Over recent years, deep learning technologies have achieved significant progress in the field of computer vision, and many tampering detection models have started to leverage these technologies. However, most existing models require pre-training on large datasets and exhibit relatively weak robustness and generalizability. To address these issues, this

收稿日期:2022-11-15;在线发布日期:2023-05-08. 本课题得到国家自然科学基金委员会基金(U19B2022,62272314,U22B2047)、广东省自然科学基金杰出青年项目(2019B151502001)、深圳市基础研究项目(JCYJ20200109105008228)资助. 胡林辉,硕士研究生,主要研究领域为图像篡改检测、深度学习. E-mail:2110276166@email.szu.edu.cn. 陈保营,硕士,主要研究领域为图像取证、视频取证、深度学习. 谭舜泉(通信作者),博士,副教授,硕士生导师,中国计算机学会(CCF)会员,主要研究领域为多媒体安全、多媒体取证、机器学习. E-mail:tansq@szu.edu.cn. 李斌,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为多媒体取证、图像处理、深度学习.

study employs the Convnext model, a high-performing pure convolutional neural network in the computer vision domain, as the backbone network. Furthermore, the study utilizes the Unified Perceptual Parsing network (Upernet) to extract multi-scale features from images, constructing a tampering detection and localization model based on the Convnext-Upernet framework. Building on this foundation, the study further applies a self-supervised data augmentation technique to amplify tampering traces in images and enhances the accuracy of tampering detection and localization by employing an image classification loss function unrelated to the tampering detection and localization task. This novel approach allows the model to focus on detecting subtle tampering artifacts without being influenced by unrelated factors present in the input images. Large-scale experimental results on mainstream tampering detection and localization datasets, demonstrate that the proposed model exhibits efficient and precise tampering detection and localization capabilities. Compared with the state-of-the-art cross-database performance of MVSS-Net++, the proposed model improves detection and localization performance by 14.4%, achieves comprehensive generalizability enhancement, and demonstrates strong robustness against common post-processing operations, such as JPEG compression and resizing. In conclusion, this study presents a novel Convnext-Upernet-based tampering detection and localization model that addresses the limitations of current models by incorporating self-supervised data augmentation and an image classification loss function. The proposed model outperforms existing state-of-the-art methods, providing a promising solution for mitigating the negative impacts of manipulated images in the digital era. Future research directions include exploring the integration of other advanced deep learning techniques to further improve the model's performance and developing real-time detection systems that can be deployed in various applications, such as social media platforms and news agencies.

**Keywords** image forgery detection; deep learning; convolution network; unified perceptual parsing network; self-supervised data augmentation

## 1 引 言

随着图像处理技术和软件行业的不断发展,人们对图像进行处理的成本变得越来越低,操作变得越来越便捷.在大数据时代,图像的传播也越来越广泛.然而,这也给正确检测图像的真实性带来了挑战.近年来,国内外关于恶意篡改图像的报道层出不穷<sup>[1-2]</sup>.如果不能有效地检测恶意篡改图像并精准地定位其篡改区域,会对社会稳定和国家安全造成很大的负面影响.

本文旨在探讨三种常见的图像篡改操作:复制移动(copy-move)、拼接(splice)和移除(removal).复制移动是指将图像的一部分复制粘贴到图像的另一位置,而拼接是指将一张图像的一部分复制并粘贴到另一图像中.移除操作则是通过算法擦除图像中的某个物体,并还原其背景区域.本文通过图1展示了三种篡改方式的实例,其中从上至下分别为真

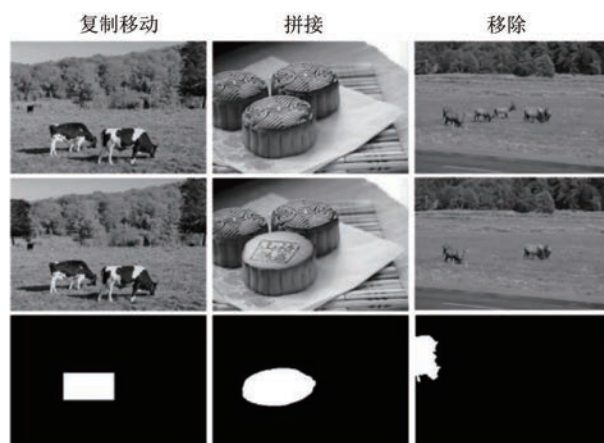


图1 篡改图像实例

实图像、篡改图像以及相应的篡改区域掩码.其中,黑色表示真实图像区域,而白色则表示篡改区域.

目前针对图像篡改检测的方法主要分为基于手工特征的传统检测方法和基于深度学习的检测方法.

在图像篡改检测工作初期,大多数工作都采用基于手工特征的传统检测方法.传统检测方法从篡改手段分为基于重叠块<sup>[3-4]</sup>、基于特征点<sup>[5]</sup>、基于图像属性<sup>[6]</sup>、基于设备属性<sup>[7]</sup>和基于压缩属性<sup>[8-9]</sup>.

但是,传统方法大多只针对单一的篡改方式进行检测,泛化性能较差.这些方法中的绝大部分只能检测图像是否被篡改,无法用于篡改区域定位.随着深度学习方法不断发展,近年来出现了许多基于深度学习的图像篡改检测定位方法.

自从深度学习被应用于图像篡改检测任务,大多数模型都是基于可学习的卷积神经网络(CNN)建立图像篡改检测模型.Rao等人<sup>[10]</sup>首次将卷积神经网络应用于图像篡改检测.与传统方法类似,早期的深度学习方法针对不同的篡改方式设计了不同的模型.

针对复制移动的篡改检测,Wu等人<sup>[11]</sup>提出了一种端到端的模型BusterNet.该模型不仅能够检测图像中的复制移动操作,而且能够精确定位复制源区域和目标区域.

针对拼接的篡改检测,Salloum等人<sup>[12]</sup>提出了一种基于完全卷积网络(FCN)的定位拼接图像模型.该模型具有两个分支,一个用于检测篡改区域,另一个用于检测篡改区域的边界.最后将篡改区域边界进行填充,然后与篡改检测区域取交集,即为最终结果.

针对移除的篡改检测,Li等人<sup>[13]</sup>提出了一种基于图像残差的高通全卷积网络(HP-FCN).该模型根据修复图像在残差域中的转移概率远低于未修复图像的特点,使用四个残余块(Residual Blocks)串联起来,形成了一个特征提取器.最后通过上采样将特征图放大到原尺寸.

在此基础上,研究人员还设计了对任意篡改方式通用的检测方法.Bappy等人<sup>[14]</sup>提出了J-LSTM,它使用长短期记忆模型(LSTM)框架,并利用篡改区域边界差异构建了篡改检测定位框架.Bappy等人<sup>[15]</sup>在此基础上进一步提出了H-LSTM,该模型利用重采样特征、LSTM和编码器-解码器网络的框架来进行篡改图像的定位.Zhuang等人<sup>[16]</sup>设计了一种全卷积结构,并采用稠密连接和空洞卷积来实现更好的检测定位性能.Wu等人<sup>[17]</sup>提出了一个不需要预处理和后处理的全卷积网络ManTra-Net,该网络包含两个子网络:操作痕迹特征提取器(MTE)和局部异常检测网络(LADN),将篡改检测任务当作局部异常检测任务,使用Z-score特征捕获局部异常,

使用LSTM进行评估.然而,在ManTra-Net中,模型仅对不同尺寸的特征图进行建模,而不会对图像块之间的空间关系进行建模.因此,Hu等人<sup>[18]</sup>改进了ManTra-Net,提出了SPAN(Spatial Pyramid Attention Network).SPAN由三个部分组成:特征提取器、空间金字塔注意模块和预测模块.使用的特征提取器与ManTra-Net相同,空间金字塔注意模块用于建立像素级别的多尺度空间关系,预测模块使用卷积网络判断像素是否被篡改.由于SPAN还未能充分利用空间相关性,Liu等人<sup>[19]</sup>提出了PSCC-Net,这是一个渐进式预测的模型.PSCC-Net利用密集的交叉连接的不同尺寸的特征从粗到细产生预测掩码.得益于这种模式,PSCC-Net具有非常不错的鲁棒性,因此也是本文比较模型鲁棒性的主要对象.

近年来,手工特征与深度学习相结合的篡改检测网络进一步提高了模型性能.Zhou等人<sup>[20]</sup>提出了双流RGB-N网络,双流分别为从RGB图像提取特征的RGB流和利用隐写分析富模型<sup>[21]</sup>滤波提取噪声特征来发现真实区域和篡改区域之间的噪声一致性的噪声流.Chen等人<sup>[22]</sup>通过多视角特征学习和多尺度监督提出了MVSS-Net++,它由噪声分支和边界分支组成,噪声分支用于学习语义无关特征,从而获得更加通用的特征;边缘分支用于学习篡改区域与真实区域在边界处的不一致性.它是目前最先进,特别是在跨数据集能力最优秀的模型,同时也是本文的主要比较对象.

综上所述,虽然基于深度学习的篡改检测方法在近年迅速发展,但仍然存在一些问题.基于深度学习的篡改检测方法虽然在单一数据集上取得了较好的效果,但模型的鲁棒性和泛化能力较差.为了解决这一问题,目前对于提升篡改后图像后处理的鲁棒性和跨数据集的泛化性的方案,还是先采用较大数据集进行预训练,然后在特定的数据集上进行微调.而创建大规模数据集是一项耗费人力和物力的工作.针对这个问题,本文利用Upernet<sup>[23]</sup>结构进行篡改检测,引入图片分类损失函数提高模型效率,采用自监督数据增强来提高模型的鲁棒性和泛化性.

本文的主要贡献有:

(1)针对目前在大规模数据集预训练的方法训练时间长,成本高的缺点,我们提出了自监督的数据增强.使得训练图像能自动插入多种篡改痕迹,增加了数据的多样性.这一方法无需进行预训练-微调,可以直接获得很好的效果,从而大大提高了模型

的性能;

(2)针对篡改检测定位任务中常用的像素级分类损失函数,本文添加了计算机视觉领域的图像分类损失函数作为辅助损失函数,使得网络更好地收敛,进一步提升了模型性能;

(3)针对图像篡改检测任务,我们将计算视觉领域中的 Convnext<sup>[24]</sup>和 Upernet<sup>[23]</sup>相结合,构建了基于 Convnext-Upernet 的图像篡改检测模型.相较于过去的工作中大多只关心是否正确定位篡改区域,而忽略了对载体图像的检测,本文的模型不仅能够精确定位篡改区域,还能正确检测载体图像.大量实验证明了本文模型的优越性.

本文在第2节介绍本文使用的模型;第3节介

绍实验过程和结果;最后第4节总结全文.

## 2 模型结构

本文的任务是对一张可疑的图像进行篡改检测,判断该图像是否被篡改,并定位篡改区域.我们采用的 Convnext-Upernet 模型结构如图2所示,主要由特征金字塔和金字塔池化模块构成. Convnext-Upernet 在特征金字塔网络的主干网络的最后一层添加了一个金字塔池化模块,然后将其输入至特征金字塔网络自顶向下的分支.图中灰色区域代表特征金字塔网络,金字塔池化模块连接与  $C_5$  特征图后,两者具体的结构将在本节(2.2节)进行介绍.

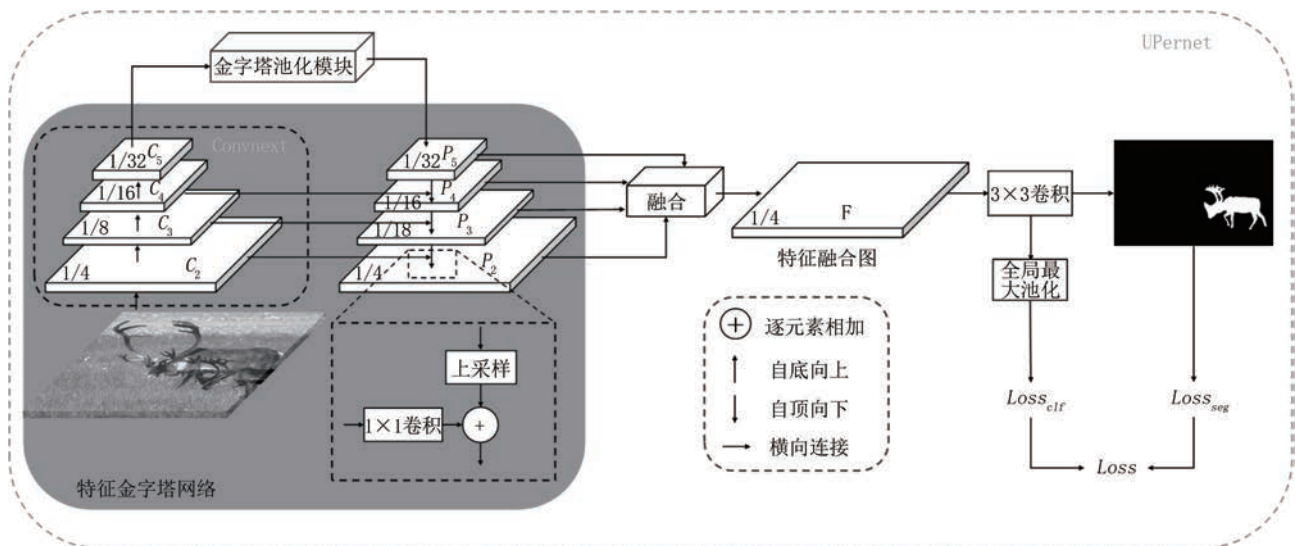


图2 Convnext-Upernet整体结构

设输入图像的尺寸为  $H \times W \times C$ , 表示输入图像的长、宽和通道数. 输入图像首先经过 Convnext<sup>[24]</sup>提取特征, 得到4个不同尺寸的特征图 ( $C_2, C_3, C_4, C_5$ ), 四个特征图的大小分别是原特征图的  $1/4, 1/8, 1/16$  以及  $1/32$ ,  $C_5$  特征图在金字塔池化模块中融合了4种经过不同尺寸的全局平均池化层得到特征图  $P_5$ , 将  $P_5$  进行3次上采样并在每次上采样时和 Convnext 得到对应尺寸的特征图进行融合得到  $P_4, P_3, P_2$ . 将  $P_5, P_4, P_3, P_2$  通过双线性插值放大到原尺寸然后融合. 最后由  $3 \times 3$  的卷积对每个像素进行分类并通过双线性插值将尺寸还原到原尺寸  $H \times W$  得到预测结果.

### 2.1 Convnext

Convnext是目前性能较好的卷积神经网络,它以 ResNet50<sup>[25]</sup>作为基础,借鉴 Swin Transformer<sup>[26]</sup>

的设计理念. 在 ImageNet<sup>[27]</sup>数据集上进行预训练,在各项任务上都超过了 Swin Transformer.

本文使用 Convnext 系列中的 Convnext-B 作为主干网络对图像进行特征提取,它在每个阶段中的通道数  $C$  和块数  $B$  分别为:  $C=(128, 256, 512, 1024)$ ,  $B=(3, 3, 27, 3)$ . Convnext-B 内部模块如图3所示. 其中  $x \in (1, 2, 4, 8)$  代表着 Convnext-B 中不同维度的内部模块, LN 为层归一化 (Layer normalization), GELU (Gaussian Error Linear Units) 为激活函数. 以我们使用的 Convnext-B 主干网络为例,主干网络共四层,每层分别具有  $(3, 3, 27, 3)$  块图3所示的内部模块,每层最后输出的维度分别为  $(128, 256, 512, 1024)$ .

### 2.2 Upernet

Upernet 主要由金字塔池化模块 (PPM) 和特征金字塔网络 (FPN) 两部分组成. 输入图像经过

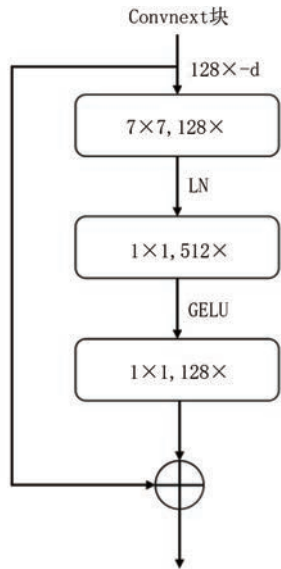


图3 Convnext-B 内部模块

Convnext 提取特征得到特征图  $C_5$ ,  $C_5$  进入金字塔池化模块后首先经过 4 种不同尺度的平均池化层, 然后再通过双线性插值恢复到原特征图尺寸, 最终全

部拼接到一起得到  $P_5$  送入特征金字塔网络. 普通的 Upernet<sup>[23]</sup> 在特征融合图中连接了许多分支用于不同的任务, 在我们的 Convnext-Upernet 中只使用了其中一个分支进行篡改检测.

在深度卷积网络中, 网络实际的感受野比理论上要小很多, 为了解决这一问题, 在主干网络最后一层加入金字塔池化模块. 金字塔池化模块获取不同子区域表示, 然后进行上采样和拼接, 将多尺度特征进行融合从而达到获取全局信息的目的. 金字塔池化模块的内部结构如图 4 所示.

图 4 中 4 种不同大小的特征图表示着经过不同大小的平均池化所得到的结果. 其内部所涉及的操作如公式 1 所示.

$$P_5 = Bottleneck(Concat(Pool_i(x))), i \in [1, 4] \quad (1)$$

其中 *Bottleneck* 代表瓶颈层, 由  $3 \times 3$  卷积, 正则化, Relu 激活函数以及置弃层(Dropout)构成, *Concat* 为拼接操作,  $Pool_i$  代表不同尺寸的均值池化层,  $x$  为输入图像.

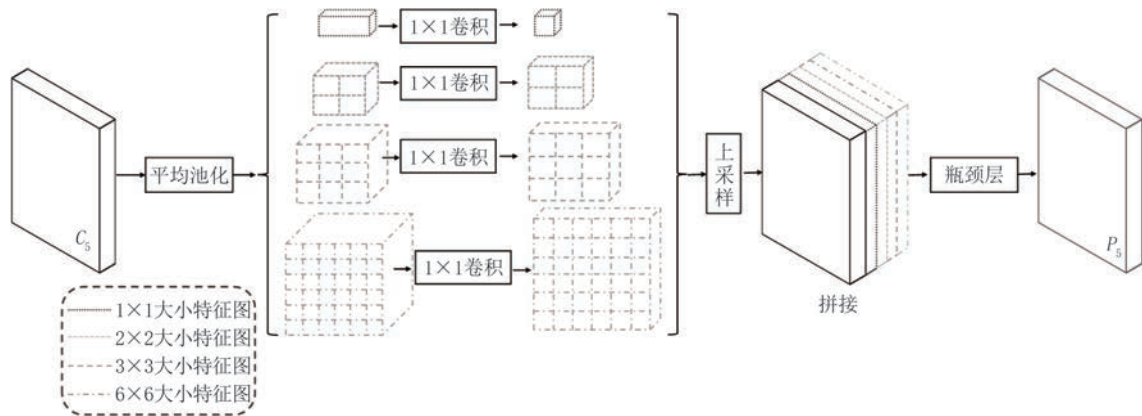


图4 金字塔池化模块结构(不同大小的线条表示经过不同尺寸的池化层得到的特征图)

经过金字塔池化模块后将特征图送入特征金字塔网络, 它由 3 部分组成, 分别是自底向上的线路、自顶向下的线路和横向连接. 自底向上的线路在本文中其实就是 Convnext 主干网络特征提取的过程, 在特征提取过程中分别得到了  $C_2, C_3, C_4, C_5$  共 4 个不同尺寸的特征图. 特征金字塔网络中的自顶向下和横向连接操作过程如图 2 中的虚线框所示. 在自顶向下中通过上采样将小尺寸的特征图一步步放大到特征提取时对应的尺寸; 于此同时进行横向连接, 横向连接将经过上采样后和当前分辨率一致的特征, 通过相加的方式进行融合得到  $P_2, P_3, P_4, P_5$ . 这样融合了多尺度的特征图  $F$  具有丰富的特征以供模型进行篡改检测. 特征金字塔网络中将 4 个不同尺

度的特征图融合的相关公式如公式 2 和公式 3 所示.

$$F = Conv(Concat(P_i)), i \in [2, 5] \quad (2)$$

$$P_i = Conv_{1 \times 1}(C_i) + Up(P_{i+1}), i \in [2, 4] \quad (3)$$

其中公式(2)中, *Conv* 表示图 2 中的融合块, 由  $3 \times 3$  卷积、批量正则化和 Relu 激活函数构成. 融合块将 4 个特征图  $P_2, P_3, P_4, P_5$  拼接后再进行降维并将结果缩放到输入图像大小. 公式(3)中,  $Conv_{1 \times 1}$  为  $1 \times 1$  卷积, *Up* 表示为采用双线性插值的上采样操作.

至此完成了一次篡改检测定位过程, 对  $F$  进行二值化处理即可得到最后模型预测的篡改区域. Convnext 可以很好地提取图像特征, 配合 Upernet 中的横向连接操作和金字塔池化模块能够更好的利

用 Convnext 提取的特征. 我们在消融实验(3.5节)中进一步验证了这一点. 本文模型在进行篡改检测定位时取得了很好的性能.

### 2.3 自监督数据增强

在数据增强方面, 基础的数据增强包括随机上下和水平翻转, 随机 JPEG 压缩和随机中值模糊. 此外, 我们还设计了两种自监督数据增强技术: 自监督复制移动数据增强和自监督拼接数据增强, 以增强每张篡改图像的多样性. 与以往的方法不同的是, 我们不需要在大规模的数据集上进行预训练, 然后再在各个数据集上进行微调. 相反, 我们希望在自监督数据增强的帮助下, 使模型仅在较小的数据集上进行训练, 就能够达到需要进行预训练模型的性能, 如 ManTra-Net<sup>[17]</sup>和 SPAN<sup>[18]</sup>等模型.

自监督数据增强是一种利用已有数据, 根据不同任务通过对数据进行变化或操作从而生成新样本的方法. 新样本可以使模型对数据的变化更具有适应性, 提高模型的性能. 在自监督数据增强中, 我们不需要额外的标注数据, 而是利用目前的数据来生成合成样本, 这些样本可以用于训练模型, 从而提高模型的泛化能力和鲁棒性. 在不同的研究领域中采取不同的方式实现自监督数据增强方法. 例如, Zhang 等人<sup>[28]</sup>在情绪识别中通过对抗网络合成逼真的脑信号. Chen 等人<sup>[29]</sup>在行人重识别中通过将行人的上半身和下半身随机拼接获得更多的数据.

#### 算法 1. 复制移动自监督数据增强

输入: 一个篡改图像  $I$ , 篡改图像对应的篡改区域掩码  $Y_{gt}$  和代表是否进行随机缩放和裁剪的布尔值  $Is\_random$

输出: 新的复制移动篡改图像  $I'$  和对应篡改区域掩码  $Y'_{gt}$

1. 使用能够包围住篡改区域的最小矩形从  $I$  中裁剪篡改区域  $R_r$  和其对应篡改区域  $M_r$ , 相似的, 随机从  $I$  中裁剪出和  $R_r$  相同尺寸的  $H, W$  的矩形区域  $R_r$  和其对应篡改区域  $M_r$ ;

2. 如果  $Is\_random$  为 True 然后: 随机选择缩放或者裁剪, 如果选择了缩放则将  $R_r, M_r$  按  $(0.5, 1)$  随机比例缩放, 如果选择了裁剪则将  $R_r, M_r$  从左上角裁剪到大小:

$$H' \in [2, H], W' \in [2, W];$$

3. 从 0 到 1 区间取随机数:  $p_1 = random(0, 1)$ ;

4. 如果  $p_1 < 0.5$  然后: 把裁剪出的  $R_r$  按照下面公式粘贴到  $R_r$  上形成新的篡改图像  $I'$  和其对应篡改区域  $Y'_{gt}$ :

$$R_r \times M_r + R_r \times (1 - M_r) \rightarrow I', M_r \cup M_r \rightarrow Y'_{gt};$$

5. 否则: 把裁剪出的  $R_r$  按照下面公式粘贴到  $R_r$  上形成新的篡改图像  $I'$  和其对应篡改区域  $Y'_{gt}$ :

$$R_r \times M_r + R_r \times (1 - M_r) \rightarrow I';$$

6. 假设结束;

7. 返回新的复制移动篡改图像和其对应篡改区域  $I'$  和  $Y'_{gt}$ ;

复制移动自监督数据增强的算法流程如算法 1 所示. 算法首先通过原图像的篡改区域掩码获得篡改区域图像, 然后将篡改区域图像随机缩放或裁剪, 最后将篡改区域随机粘贴到原图像得到新的篡改图像. 在复制移动自监督数据增强中, 为了使复制移动篡改内容不局限与其篡改区域, 我们增加了一个范围在 0 至 1 之间均匀分布的随机数  $p_1$ , 若  $p_1 > 0.5$  则将裁剪的区域粘贴到原篡改区域形成新的篡改图像. 值得注意的是, 只有当参数  $Is\_random$  为 True 时, 才会对篡改区域图像进行随机缩放或裁剪, 此时我们将自监督数据增强称作随机自监督数据增强. 拼接自监督数据增强同理. 这么做的原因是为了进一步提高数据的多样性.

拼接自监督数据增强的算法流程如算法 2 所示. 首先从拼接图像摘取出其对应的篡改区域图像, 其次将篡改区域图像随机缩放, 最后将其随机粘贴到原图像获得新的篡改图形. 在拼接自监督数据增强中, 为了获得更加自然的拼接图像, 我们增加了一个范围在 0 至 1 之间均匀分布的随机数  $p_2$ , 若  $p_2 < 0.2$  则在拼接时进行泊松融合<sup>①</sup>形成新的篡改图像.

#### 算法 2. 拼接自监督数据增强

输入: 两个篡改图像  $I_1, I_2$ , 篡改图像对应篡改区域掩码  $Y_1, Y_2$  和代表是否进行随机缩放的布尔值  $Is\_random$

输出: 新的拼接篡改图像  $I'$  和对应篡改区域  $Y'_{gt}$

1. 使用能够包围住篡改区域的最小矩形从  $I_1$  中裁剪篡改区域  $R_1$  和其对应篡改区域  $M_1$ , 相似的, 随机从  $I_2$  中裁剪出和  $R_1$  相同尺寸的  $H, W$  的矩形区域  $R_2$  和其对应篡改区域  $M_2$ ;

2. 如果  $Is\_random$  为 True 然后: 进行随机缩放, 将  $R_2, M_2$  按  $(0.5, 1)$  随机比例缩放;

3. 把裁剪出的  $R_1$  按照下面公式粘贴到  $R_2$  上形成新的篡改图像  $I'$  和其对应篡改区域  $Y'_{gt}$ ;

4. 从 0 到 1 区间取随机数:  $p_2 = random(0, 1)$ ;

5. 如果  $p_2 < 0.2$  然后: 对  $I'$  和  $I_2$  进行泊松融合;

6. 假设结束;

7. 返回新的复制移动篡改图像和其对应篡改区域  $I'$  和  $Y'_{gt}$ ;

使用自监督数据增强能增加篡改数据的多样性, 使得一张篡改图像中有多种篡改方式, 模型能从

① 泊松融合采用 OpenCV 中的 seamlessClone 方法进行

一张图像学习到更多的特征. 复制移动自监督数据增强样例如图 5 所示, 拼接自监督数据增强图像样例如图 6 所示.

在图 5 中, 情形“复制移动 1”为算法 1 中  $p > 0.5$  的情况, 即将一部分背景复制到篡改区域, 情形“复制移动 2”为算法 1 中  $p < 0.5$  的情况, 即将原篡改区域再次复制. 图 5 原篡改图像(左图, 情形“篡改图像”)为拼接篡改图像, 经过复制移动自监督数据增强后篡改图像(右图, 情形“复制移动 2”)同时具有复制移动篡改图像和拼接篡改图像的特征.

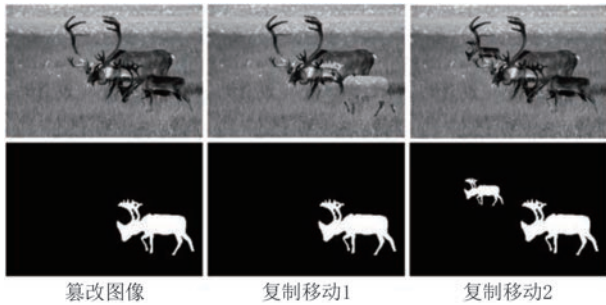


图 5 复制移动自监督数据增强样例

图 6 为两张篡改图像经过拼接自监督数据增强后得到的结果. 图像(c)为图像(b)拼接到图像(a)中生成的, 同理图像(d)为图像(a)拼接到图像(b)中所生成的. 图像经过两种自监督数据增强处理后, 能够生成更加丰富的篡改特征.

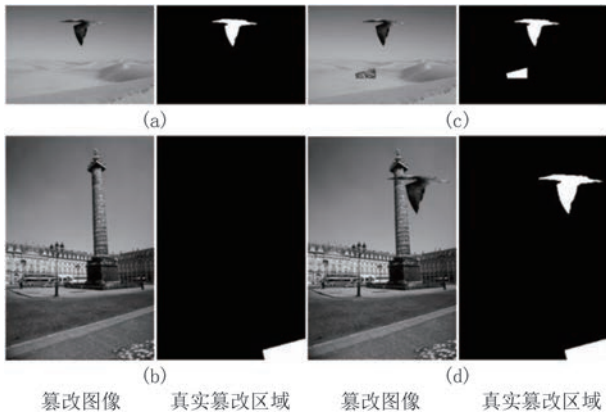


图 6 拼接自监督数据增强样例

## 2.4 损失函数

过去的工作大多直接使用二分类交叉熵损失函数( $BCELoss$ ,  $L_{BCE}$ , 如公式 4 所示)对模型进行训练. 然而, 当篡改区域占整张图像比例较小时, 损失函数中  $y=0$  的成分会占据主导, 导致模型更加偏向非篡改区域. 为了解决这个问题, 我们在  $BCELoss$  的基础上加入了用于图像分割领域的  $DiceLoss$  ( $L_{Dice}$ ,

如公式 5 所示)组成联合损失函数  $BCEDiceLoss$  ( $Loss_{seg}$ , 如公式 6 所示).  $DiceLoss$  来自于  $dice$  系数.  $dice$  系数是一种评估两个样本相似性的系数, 取值在  $[0, 1]$  之间, 值越大表示两样本相似度越高.  $dice$  系数分子部分代表了预测结果和篡改区域掩码的交集, 分母部分为预测结果和篡改区域掩码中篡改像素的数量.  $DiceLoss$  通常用于解决正负样本严重不平衡的现象, 可以很好地解决篡改图像中篡改区域较小的问题.  $DiceLoss$  与  $BCELoss$  结合后可以更好地解决由于篡改区域较小导致无法正常训练的问题.  $Loss_{seg}$  代表整张图片中对每个像素进行分类的损失函数, 在损失函数公式中(公式 4-公式 6),  $\lambda = 0.1$ ,  $G(x)$  表示 Convnext-Upernet 模型,  $X_{i,j}$  表示输入图像第  $i$  行第  $j$  列对应像素,  $Y_{i,j}$  表示对应位置像素的掩码值, 我们使用  $Loss_{seg}$  作为训练模型的基础损失函数.

$$L_{BCE}(x) = -\frac{1}{W \times H} \sum_{i=1}^H \sum_{j=1}^W (Y_{i,j} \cdot \log G(X_{i,j}) + (1 - Y_{i,j}) \cdot \log(1 - G(X_{i,j}))) \quad (4)$$

$$L_{Dice}(x) = 1 - dice \quad (5)$$

$$dice = \frac{2 \times \sum_{i=1}^H \sum_{j=1}^W (G(X_{i,j}) \times Y_{i,j})}{\sum_{i=1}^H \sum_{j=1}^W G^2(X_{i,j}) + \sum_{i=1}^H \sum_{j=1}^W (Y_{i,j}^2)} \quad (5)$$

$$Loss_{seg} = L_{BCE} + \lambda L_{Dice} \quad (6)$$

在  $Loss_{seg}$  的基础上, 我们还加入了  $Loss_{clf}$  作为辅助损失函数,  $Loss_{clf}$  代表着对图像进行分类的损失函数, 即输入图像是否为篡改图像的二分类交叉熵损失函数. 在经过 Convnext-Upernet 的特征金字塔网络得到特征融合图后, 进一步使用卷积对每个像素进行分类并产生两个分支, 一个分支计算  $Loss_{seg}$ , 另一分支经过最大池化层得到整幅图是否为篡改图像的概率, 采用二分类交叉熵损失函数, 计算图片分类损失函数  $Loss_{clf}$  如公式 7 所示, 其中  $X$  为输入图像,  $Y$  为图像标签,  $Y=1$  表示为篡改图像,  $Y=0$  表示为载体图像,  $GMP$  (Global Max Pooling) 表示为全局最大池化,  $Loss_{seg}$  与  $Loss_{clf}$  两者相加作为最终的损失如公式 8 所示.

$$Loss_{clf} = -Y \cdot \log(GMP(G(X))) + (1 - Y) \cdot \log(1 - GMP(G(X))) \quad (7)$$

$$Loss = Loss_{seg} + Loss_{clf} \quad (8)$$

我们使  $Loss_{seg}$  和  $Loss_{clf}$  权重相等, 保证图像篡改检测和图片分类任务平等. 在消融实验(3.5节)中我们通过实验证明了图片分类损失函数的实用性.

### 3 实验及结果

本文在目前5个主流的篡改检测数据集对Convnext-Upernet性能进行验证,这5个数据集分别为:CASIA<sup>[30]</sup>、Columbia<sup>[31]</sup>、COVERAGE<sup>[32]</sup>以及NIST16<sup>[33]</sup>和IMD2020<sup>[34]</sup>.主要测试了Convnext-Upernet的跨库性能,库内性能,模型鲁棒性.通过消融实验,探究了自监督数据增强、损失函数以及模型中不同组件对性能的影响.为了更好地配合图像分类损失函数,在一些实验设置中还加入了载体图像进行训练.

#### 3.1 数据集和实验设置

##### 3.1.1 数据集

本次实验主要使用了5个主流篡改检测数据集,下面依次对这些数据集进行介绍:

(1) CASIA:该数据集主要包含复制移动和拼接篡改图像,有CASIAv1和CASIAv2两个版本,我们使用CASIAv2中的篡改图像进行训练和验证,CASIAv1中的篡改图像进行测试.在消融实验中,将CASIAv2中的3745张载体图像加入训练和验证.

(2) Columbia:该数据集集中于未经压缩的拼接图像,数据集较小,共有183张高分辨率篡改图像.

(3) COVERAGE:该数据集是一个复制移动篡改图像数据集,包含100张篡改图像.

(4) NIST16:该数据集共包含复制移动,拼接,移除三种不同的篡改方式,图像分辨率高,数据集的大小为564.

(5) IMD2020:该数据集是近期新出现的篡改检测数据集,目前针对该数据集进行跨库测试的结果较少.它从2322种相机型号收集了35000张载体图像,并通过GAN或者Inpainting等方法生成了35000张图像以及2000张手工篡改图像.由于这35000张图像并没有提供真实篡改区域掩码,我们所使用训练及验证的数据是这手工生成并提供篡改区域掩码的2000张图像.

##### 3.1.2 实验设置

在训练时,我们使用了CASIA数据集中原始尺寸的篡改图像.对于其余四个数据集,我们将图像尺寸调整为 $512 \times 512$ .本文的实验环境为:Python=3.7、Pytorch=1.9.0、Torchvision=0.10.0、使用的显卡为Tesla V100、使用的优化器为Adam、学习率

(learning rate)为0.0001、我们一共训练了100个迭代次数(epoch)、批次大小(batchsize)为16.

当前针对篡改检测的数据集数量普遍偏少,因此我们采用在ImageNet预训练的Convnext-B对上述数据集进行训练,各数据集在训练时所包含的样本数如表1所示.由于COVERAGE数据集数量较少,没有设置验证集.除CASIA数据集外,其余数据集的训练集,验证集和测试集的比例为0.7:0.05:0.25.

表1 各数据集训练样本数

数据集	训练集	验证集	测试集
CASIA	4597	511	920
Columbia	729	9	45
COVERAGE	75	-	25
NIST16	394	28	142
IMD2020	1205	302	503

#### 3.2 评价指标

本文采用像素级曲线下面积(Area under ROC curve, AUC)、F1分数(F1-score)、交并比(IOUS)和漏检率(FNR)对模型进行评估,对每个像素判断是否经过篡改,其中F1分数的计算方式如公式9所示.

$$F1 = \frac{2 \times TP}{2 \times TP + FN + FP} \quad (9)$$

$$IOU = \frac{|\hat{y} \cap y|}{|\hat{y} \cup y|} \quad (10)$$

$$FNR = \frac{FN}{TP + FN} \quad (11)$$

其中,TP表示模型预测正确的篡改像素点数目,FP表示模型预测错误的篡改像素点数目,FN表示模型预测错误的载体像素点数目, $\hat{y}$ 与y分别表示预测掩码和真实掩码.

#### 3.3 实验结果

##### 3.3.1 跨数据集能力测试

我们首先对模型的跨库能力进行测试,我们在相同的设置下比较了模型的跨库能力.其中,ManTra-Net<sup>[17]</sup>为论文中在自制数据集进行预训练后所报告的数据.由于我们无法完全复现ManTra-Net论文中所报道出的数据,我们利用GitHub上的模型并采取和本文模型相同的实验设置进行实验.为了与ManTra-Net论文结果进行区分,我们将上述实验记作Mantra-Net\*.MVSS-Net++<sup>[22]</sup>为在CASIAv2数据集上进行训练的实验结果,为了实验



公平,我们在训练MVSS-Net++时加入了本文提出的自监督数据增强.模型在CASIAv2训练后在各个数据集上的像素级性能如表2所示.在同一个数据集中性能最好的用加粗表示,“-”代表在该论文

中并没有对该指标进行测试或没有在该数据集上进行测试.在测试中,对于小数据集如Columbia、COVERAGE和NIST我们使用全部的数据进行测试.其他数据集按表1的数量进行测试.

表2 Convnext-Upernet在CASIAv2训练后在不同数据集上的像素级性能(AUC/F1/IOU/FNR)%

方法	CASIAv1	Columbia	COVERAGE
ManTra-Net <sup>[17]</sup>	81.70/-/-/-	82.40/-/-/-	81.90/-/-/-
ManTra-Net*	80.46/34.96/24.27/48.95	72.08/46.83/32.34/49.31	69.90/26.10/15.85/55.24
MVSS-Net++ <sup>[22]</sup>	89.07/59.37/50.12/27.96	83.97/66.60/56.37/35.64	<b>87.43/48.57/38.52/38.07</b>
本文模型	<b>98.60/84.69/78.21/13.76</b>	<b>95.70/88.24/83.57/10.55</b>	86.43/38.57/31.9/59.02
方法	NIST16		IMD2020
ManTra-Net <sup>[17]</sup>	79.5/-/-/-		-/-/-/-
ManTra-Net*	71.33/18.05/10.96/55.48		76.04/22.75/14.34/ <b>41.83</b>
MVSS-Net++ <sup>[22]</sup>	83.19/41.22/32.15/ <b>44.43</b>		81.42/34.01/25.12/49.88
本文模型	<b>87.30/45.06/36.77/45.41</b>		<b>90.35/45.62/37.04/45.05</b>

从表2可以看出,在CASIAv2训练后我们的模型在5个公开数据集上都有着不错的性能,在多个数据集上的AUC、F1、IOU以及FNR性能上都超过了ManTra-Net<sup>[17]</sup>和MVSS-Net++<sup>[22]</sup>.即使在有着三种不同篡改方式的NIST16数据集上也有着明显的提升,证明了我们的模型有着较好的泛化性,跨库性能同样十分不错.虽然在NIST16数据集上的误检率比MVSS-Net++略低,但是在其他三个指标上均超过了MVSS-Net++.在与Mantra-Net\*和MVSS-Net++的对比中可以证明本文模型的优越性.

值得注意的是,相较于ManTra-Net<sup>[17]</sup>等方法,本文只在CASIAv2小数据集上进行训练,且无需微调即取得了较好的效果.这种方式降低了人工制作数据集的难度和训练时间.与同样在CASIAv2上训练的MVSS-Net++<sup>[22]</sup>相比,本文模型表现更出色.

### 3.3.2 数据集内部测试

我们接下来在4个公开数据集内进行训练和测试,其训练集和测试集按表1分配.Convnext-Upernet与其他模型在不同数据集上的性能如表3所示.由于某些方法并没有开源,因此我们直接采用了其论文所报道的数据.

表3 Convnext-Upernet在不同数据集上的性能(AUC/F1/IOU/FNR)%

方法	CASIAv1	COVERAGE	NIST16	IMD2020
J-LSTM <sup>[14]</sup>	-	61.40/-/-/-	76.4/-/-/-	-
H-LSTM <sup>[15]</sup>	-	71.20/-/-/-	79.4/-/-/-	-
DFCN <sup>[16]</sup>	85.43/54.37/46.21/30.47	85.69/44.54/31.86/ <b>28.47</b>	89.83/48.54/39.9/38.88	89.04/45.80/35.85/34.91
RGB-N <sup>[20]</sup>	79.5/40.9/-/-	81.7/43.7/-/-	93.7/72.2/-/-	-
SPAN <sup>[18]</sup>	83.8/38.2/-/-	93.7/55.8/-/-	96.1/58.2/-/-	-
MVSS-Net++ <sup>[22]</sup>	89.07/59.37/50.15/27.96	91.27/51.44/43.09/42.17	94.00/72.86/63.91/ <b>22.32</b>	90.82/48.13/37.08/ <b>33.92</b>
PSCC-Net <sup>[19]</sup>	87.5/55.4/-/-	94.1/ <b>72.3</b> /-/-	<b>99.6</b> /81.9/-/-	-
本文模型	<b>98.60/84.69/78.21/13.76</b>	<b>95.62</b> /56.72/ <b>45.58</b> /35.48	98.82/77.75/ <b>71.75</b> /23.10	<b>93.48/54.08/44.37</b> /42.29

由表3可知,Convnext-Upernet在CASIAv1数据集上的性能都要优于其他模型.在五个评价指标中都具有一定的优势,本文模型相对于DFCN<sup>[32]</sup>、RGB-N<sup>[20]</sup>、SPAN<sup>[18]</sup>以及PSCC-Net<sup>[19]</sup>都表现得更加优秀.在COVERAGE数据集上的F1性能和NIST16数据集上的表现还是低于PSCC-Net<sup>[19]</sup>.DFCN在IMD2020上的性能分别为89.04%和45.80%,而本

文模型在IMD2020上的AUC和F1分别为93.48%和54.08%.在IOU性能上,本文模型优于其他的模型.值得注意的是,本文模型在NIST16数据集上的FNR比MVSS-Net++高,在IMD2020数据集甚至比DFCN低,这意味着MVSS-Net++和DFCN在NIST16数据集和IMD2020数据集上正确检测篡改像素的数量更多.我们认为这是因为他们将大部分

区域都定位为篡改区域所导致的,3.3.3节的模型可视化也能够很好地说明这一点.

### 3.3.3 可视化

我们对不同模型的预测结果进行可视化.图7为Convnext-Upernet与开源算法ManTra-Net<sup>[17]</sup>、

DFCN<sup>[32]</sup>、SPAN<sup>[20]</sup>以及MVSS-Net++<sup>[22]</sup>预测图像可视化对比.在可视化时,三个网络都是以0.5为阈值,得到篡改区域的概率图后,超过0.5则认为该像素被篡改,否则认为该像素未被篡改.图7中篡改区域全黑表示图像并未经过篡改.

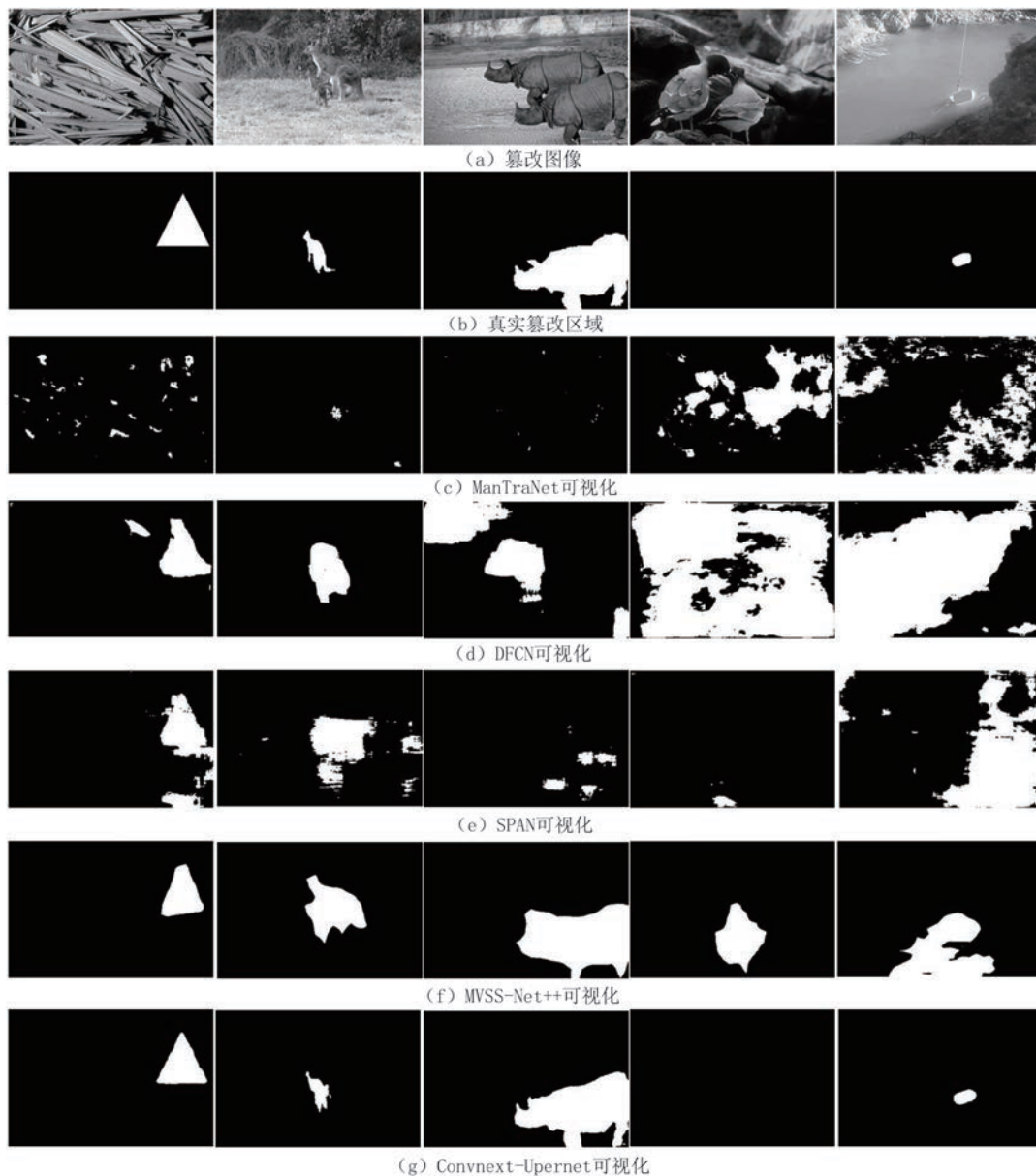


图7 不同模型预测图像可视化对比

如图7所示,Convnext-Upernet在定位篡改区域上表现得十分准确,甚至能够识别出预测载体图像中的篡改区域.相比之下,ManTra-Net和MVSS-Net++在预测未被篡改的图像时出现了误判的情况,而DFCN和MVSS-Net++的大范围预测篡改区域也解释了它们在FNR指标上的表现相对较好.这些结果进一步验证了表3中Convnext-

Upernet相较于其他模型具有更好的检测性能.

### 3.4 鲁棒性验证

为了对模型的鲁棒性进行验证,我们在NIST16和Columbia上进行各种后处理.由于NIST16和Columbia图像的分辨率较高,无后处理的情况下只是将图像缩放到 $512 \times 512$ ,后续的后处理都是基于该处理进行的.

后处理主要包括缩放<sup>①</sup>(不同尺度)、高斯模糊<sup>②</sup>( $k$ 为进行高斯模糊时核的大小)、高斯噪声<sup>③</sup>( $\sigma$ 为高斯噪声的标准差)JPEG压缩<sup>④</sup>( $q$ 为压缩质量). 我们采取与SPAN相同的鲁棒性验证设置, 因为SPAN<sup>[18]</sup>是目前在相关工作中首次提出模型鲁棒性测试任务的工作, 并且后续工作都采取和它相同的实验设置.

表4为我们的模型与SPAN<sup>[18]</sup>和PSCC-Net<sup>[19]</sup>在不同的后处理上的性能比较. 如表4所示, 我们的模型在两个数据集上都具有一定的鲁棒性. 在失真程度较小的情况下, 如缩放大小为0.78以及高斯模糊核大小为3时, Convnext-Upernet相较于SPAN<sup>[18]</sup>和PSCC-Net<sup>[19]</sup>仍然可以保持不错的性

能. 但是, 在失真程度较大的情况下, 如缩放大小为0.25、高斯模糊核大小为15、JPEG压缩质量为50时, 性能下降较明显. 这是因为PCSS-Net是一种渐进式预测的模型, 它采用并行计算不同尺度的特征, 并且从低尺度特征开始预测, 由粗到细逐步预测结果. 而本文模型是一次性直接输出结果, 在面对高强度的图像扭曲时, 这种逐步完善的模型具有一定的优势. 但在高斯噪声下, 我们的模型比PSCC-Net更优秀, 说明我们的模型对高斯噪声具有很好的抵抗性. 在Columbia进行缩放到0.78大小时我们的AUC性能反而有所上升, 但F1性能还是有所下降, 我们认为这是因为Columbia数据集中未篡改像素与篡改像素数量不平衡所导致的.

表4 在NIST16数据集和Columbia数据集上鲁棒性分析(AUC/F1)%

方法	NIST			Columbia		
	SPAN	PSCC-Net	本文模型	SPAN	PSCC-Net	本文模型
无后处理	83.95/-	85.47/-	<b>87.30/45.06</b>	93.6/-	<b>98.19/-</b>	95.70/88.24
缩放(0.78×)	83.24/-	85.29/-	<b>85.76/40.01</b>	89.99/-	93.4/-	<b>96.55/85.85</b>
缩放(0.25×)	80.32/-	<b>85.04/-</b>	79.19/32.43	69.08/-	78.41/-	<b>95.43/83.59</b>
高斯模糊( $k=3$ )	83.10/-	85.38/-	<b>86.70/45.62</b>	78.97/-	84.18/-	<b>93.92/81.72</b>
高斯模糊( $k=15$ )	79.15/-	<b>79.93/-</b>	72.45/10.47	67.70/-	<b>73.24/-</b>	67.88/3.74
高斯噪声( $\sigma=3$ )	75.17/-	78.42/-	<b>85.55/40.87</b>	75.11/-	82.64/-	<b>94.72/85.59</b>
高斯噪声( $\sigma=15$ )	67.28/-	76.65/-	<b>81.20/29.37</b>	65.8/-	74.35/-	<b>87.11/48.23</b>
JPEG压缩( $q=100$ )	83.59/-	85.4/-	<b>86.73/43.61</b>	93.32/-	<b>97.97/-</b>	94.93/86.73
JPEG压缩( $q=50$ )	80.68/-	<b>85.37/-</b>	83.15/35.68	74.62/-	89.11/-	<b>89.75/72.99</b>

### 3.5 消融实验

为了验证预训练权值、自监督数据增强和图片分类损失函数的重要性, 金字塔池化模块和特征金字塔网络中横向连接对模型性能的影响, 不同主干网络和Upernet结合的性能, 我们进行了多组对照实验. 实验使用CASIAv2数据集, 在CASIAv1上测试对比. 我们首先验证预训练权值的重要性和不同损失函数对性能的影响. 预训练权值对模型性能的影响如表5所示, 表6为不同损失函数在CASIA数据集上的对比结果. 本次实验使用 $BCELoss$ ,  $DiceLoss$ 和 $BCEDiceLoss$ 3种损失函数在CASIAv2上训练本文模型, 然后在CASIAv1上测试对比. 表7为对模型组件进行删减的对比结果, 我们主要验证

表5 预训练权值的影响

方法	CASIA	
	AUC	F1
本文模型(未使用预训练权值)	65.46	9.53
本文模型(使用预训练权值)	<b>94.44</b>	<b>70.45</b>

表6 不同损失函数的影响

损失函数	CASIA		
	AUC	F1	IOU
$BCELoss$	94.35	65.98	59.59
$DiceLoss$	93.4	64.6	58.47
$BCEDiceLoss$	<b>94.44</b>	<b>70.45</b>	<b>78.21</b>

表7 不同组件对模型性能的影响

模型	CASIA	
	AUC	F1
$FCN$	86.78	51.53
$FFN$	88.73	60.32
$PPM$	89.31	66.12
$Upernet\_no\_C_2$	90.02	67.00
$Upernet\_no\_C_3$	89.97	66.39
$Upernet\_no\_C_4$	84.98	45.16
本文模型	<b>94.44</b>	<b>70.45</b>

① 缩放使用了opencv库中的resize方法

② 高斯模糊使用了opencv库中的GaussianBlur方法

③ 高斯噪声使用numpy库中的random.normal方法生成

④ JPEG压缩使用opencv库中的imencode方法

了金字塔池化模块和特征金字塔网络中横向连接对模型的影响,我们将无金字塔池化模块和横向连接操作的模型称为FCN,将无金字塔池化模块的模型称为FFN,将无横向连接操作的模型称为PPM. 由于Convnext将特征图分为四个阶段,我们很自然地在Upernet结构中使用这四个特征图. 为了进一步证明我们模型的优越性,我们在Upernet结构中只使用其中三个阶段的特征图. 我们将弃用特征图的模型称为Upernet\_no\_C<sub>i</sub>,其中*i*∈[2,4]. 表8为不同主干网络对Upernet的影响,我们采用参数量相近的Swin-Base进行对比,由于Swin-Base对图像的限制,我们统一将图像缩放到224×224进行实验. 本次实验均未使用自监督数据增强.

表8 不同主干网络的影响

主干网络	CASIA	
	AUC	F1
Swin-Base	82.66	50.89
Convnext-B	<b>85.47</b>	<b>55.58</b>

由表5可知预训练权值对模型性能影响非常大,在没有使用预训练权值的情况下,CASIA数据集的AUC只有65.4%,F1只有10%,而使用预训练权值后AUC提高到了94.4%,F1提高到了70.5%,涨幅十分巨大. 从表6可以看出不同的损失函数对模型性能也有着不同的影响,单独使用BCELoss的性能比单独使用DiceLoss更好,使用联合损失函数BCEDiceLoss比单独的BCELoss和DiceLoss更好. 如表7所示,单独删减金字塔池化模块或者横向连接操作都会对模型性能产生巨大的影响,在两者结合时性能最优,并且深层的特征图对模型影响较大,在4个特征图同时使用时达到最好的效果. 由表8可知,Convnext与Upernet相结合能够获得更好的性能.

因此在后续消融实验中我们使用预训练权值和BCEDiceLoss作为基础实验设置. 在观察自监督数据增强和损失函数的有效性中,不同实验设置和结果如表9所示. 我们将使用基础数据增强表示为*b\_aug*,本文提出的自监督数据增强表示为*s\_aug*,本文提出的随机自监督数据增强表示为*r\_aug*,图片分类损失表示为*clf*,加入载体图像进行训练表示为*add*,"+"表示使用该组件,"-"表示未使用该组件,性能最好的使用加粗表示.

从表9中可以看出,自监督数据增强对于模型

表9 不同实验设置和结果(AUC)%

设置	设置组件					数据集
	<i>b_aug</i>	<i>s_aug</i>	<i>r_aug</i>	<i>clf</i>	<i>add</i>	CASIAv1
0	+	-	-	-	-	94.44
1	-	+	-	-	-	98.26
2	-	-	+	-	-	98.40
3	+	-	-	+	-	94.67
4	-	+	-	+	-	98.14
5	-	-	+	+	-	98.60
6	+	-	-	+	+	96.86
7	-	+	-	+	+	<b>99.05</b>
8	-	-	+	+	+	98.90

性能的提升是非常显著的. 相较于基础数据增强(设置0),自监督数据增强(设置1和设置2)使模型的性能提高了4%. 然而,相较于原损失函数(设置0和设置2),采用图片分类损失函数(设置3和设置5)的提升效果并不够明显,仅分别提高了0.27%和0.2%,并且在设置4中性能反而有所下降. 这可能是因为训练数据中图片分类种类过于单一,导致图片分类损失函数没有起到很好的作用. 因此,在使用图片分类损失函数时,为了避免训练数据全为篡改图像所造成图片损失函数中标签唯一的情况发生,我们将CASIAv2中的约3500张载体图像加入训练数据,使用CASIAv1中的所有图像进行测试. 加入载体图像后的训练数据变化如表10所示.

表10 加入载体图像后的CASIA数据集

数据集	训练集	验证集	测试集
CASIA	7968	855	1720

从表9中的设置6、7和8可以看出,在增加了测试集的数量基础上,将载体图片加入训练数据可以提高模型的性能. 添加图像分类损失(设置6)提升了2.42%. 自监督数据增强与图像分类损失函数(设置7)将性能提高到了99.05%,相较于基础数据增强提升了4.61%. 在使用了随机自监督数据增强(设置8)后,性能相较于自监督数据增强反而有所下降,我们认为这是由于随机过程所导致的. 如图8所示,我们进一步将不同设置进行可视化,以观察它们的优劣.

通过图8可以看出,在不使用自监督数据增强的情况下(设置0和设置6),通常有很大一部分篡改区域被漏检为真实区域. 然而,在加入自监督数据增强(设置2和设置7)后,在预测时有很大的提升,

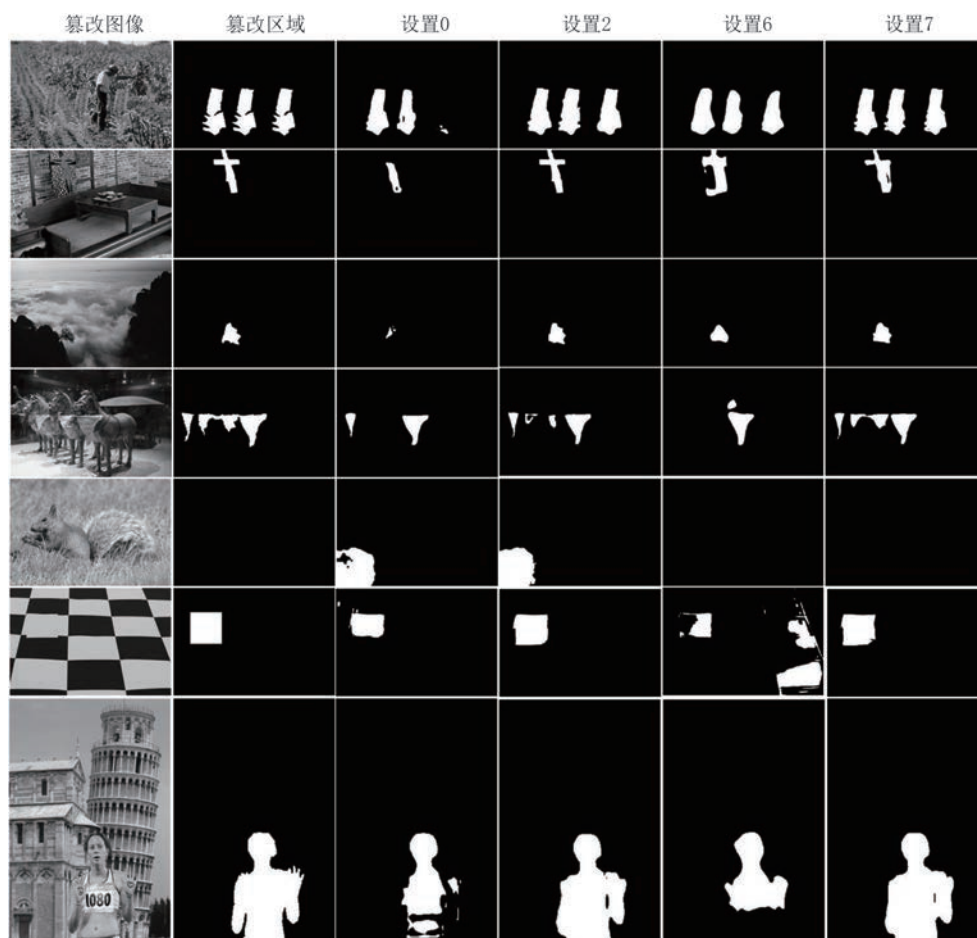


图8 不同设置下的预测结果

已经非常接近真实篡改区域了。从设置6和设置7可以看出,在训练时加入载体图像后,模型可以正确预测未经篡改的图像。

## 4 总结和展望

在信息爆炸式增长的今天,信息内容安全显得尤为重要,对图片进行篡改检测是保护信息安全的重要手段。在过去的工作中大多只考虑篡改图像,并没有考虑对载体图像的检测。我们认为在进行图像篡改检测时不仅要能正确定位篡改区域,对于未经篡改图像也要能够正确识别出来;并且模型要具有一定的跨数据集的能力。

本文结合计算机视觉中的Convnext和Upernet,构建了Convnext-Upernet,并将其应用于篡改检测定位任务中。在训练时,本文加入了自监督数据增强,以提高数据的多样性,进而提升模型性能。在基于像素的损失函数的基础上,本文加入了图像分类损失函数,以使模型更好地收敛。在多个公开数据集上测试库内性能和跨库性能,同时测试了模型的

鲁棒性。实验结果证明,在不需要对较大的数据集进行微调的前提下,本文的模型仍然能够在多个公开数据集上获得优秀的性能,并且具有不错的泛化性和鲁棒性,还能够正确检测未经篡改的图像。在下一步的工作中,本文将进一步提高模型的泛化性和鲁棒性,使其具有更强的抗攻击能力。

**作者贡献说明** 胡林辉和陈保营贡献相同。

## 参 考 文 献

- [1] Liao Mi, man in Liuzhou, Guangxi province tampered with nucleic acid test results, CCTV News, 2022(in Chinese) (廖汨, 广西柳州一男子p图篡改核酸检测结果被查, 央视新闻, 2022)
- [2] DanEvon, Did astronaut chris hadfield test the effects of marijuana in space, Snopes, 26 November, 2018
- [3] Luo W, Huang J, Qiu G. Robust detection of region-duplication forgery in digital image//Proceedings of the 18th International Conference on Pattern Recognition (ICPR). Hong Kong, China, 2006, 4: 746-749
- [4] Li G, Wu Q, Tu D, et al. A sorted neighborhood approach for

- detecting duplicated regions in image forgeries based on DWT and SVD//Proceedings of the 2007 IEEE International Conference on Multimedia and Expo(ICME). Beijing, China, 2007: 1750-1753
- [5] Amerini I, Ballan L, Caldelli R, et al. Copy-move forgery detection and localization by means of robust clustering with J-Linkage. *Signal Processing: Image Communication*, 2013, 28(6): 659-669
- [6] Dong J, Wang W, Tan T, et al. Run-length and edge statistics based approach for image splicing detection//Proceedings of the International Workshop on Digital Watermarking (IWDW). Busan, Korea, 2008: 76-87
- [7] Popescu A C, Farid H. Exposing Digital Forgeries in Color Filter Array Interpolated Images. *IEEE Transactions on Signal Processing*, 2005, 53(10): 3948-3959
- [8] Fan Z, De Queiroz R L. Identification of bitmap compression history: JPEG detection and quantizer estimation. *IEEE Transactions on Image Processing*, 2003, 12(2): 230-235
- [9] Luo W, Qu Z, Huang J, et al. A novel method for detecting cropped and recompressed image block//Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Honolulu, USA, 2007, 2: II-217-II-220
- [10] Rao Y, Ni J. A deep learning approach to detection of splicing and copy-move forgeries in images//Proceedings of the 2016 IEEE International Workshop on Information Forensics and Security(WIFS). Abu Dhabi, United Arab Emirates, 2016: 1-6
- [11] Wu Y, Abd-Elmageed W, Natarajan P. Busternet: Detecting copy-move image forgery with source/target localization//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 168-184
- [12] Salloum R, Ren Y, Kuo C C J. Image splicing localization using a multi-task fully convolutional network (MFCN). *Journal of Visual Communication and Image Representation*, 2018, 51: 201-209
- [13] Li H, Huang J. Localization of deep inpainting using high-pass fully convolutional network//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea, 2019: 8301-8310
- [14] Bappy J H, Roy-Chowdhury A K, Bunk J, et al. Exploiting spatial structure for localizing manipulated image regions//Proceedings of the IEEE International Conference on Computer Vision(ICCV). Venice, Italy, 2017: 4970-4979
- [15] Bappy J H, Simons C, Nataraj L, et al. Hybrid LSTM and Encoder-Decoder Architecture for Detection of Image Forgeries. *IEEE Transactions on Image Processing*, 2019, 28(7): 3286-3300
- [16] Zhuang P, Li H, Tan S, et al. Image tampering localization using a dense fully convolutional network. *IEEE Transactions on Information Forensics and Security*, 2021, 16: 2986-2999
- [17] Wu Y, AbdAlmageed W, Natarajan P. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA, 2019: 9543-9552
- [18] Hu X, Zhang Z, Jiang Z, et al. SPAN: Spatial pyramid attention network for image manipulation localization//Proceedings of the European Conference on Computer Vision (ECCV). Glasgow, UK, 2020: 312-328
- [19] Liu X, Liu Y, Chen J, et al. PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, doi:10.1109/TCSVT.2022.3189545
- [20] Zhou P, Han X, Morariu V I, et al. Learning rich features for image manipulation detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Salt Lake City, USA, 2018: 1053-1061
- [21] Fridrich J, Kodovsky J. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 2012, 7(3): 868-882
- [22] Dong C, Chen X, Hu R, et al. MVSS-Net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(01): 1-14
- [23] Xiao T, Liu Y, Zhou B, et al. Unified perceptual parsing for scene understanding//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 418-434
- [24] Liu Z, Mao H, Wu C Y, et al. A convnet for the 2020s//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA, 2022: 11976-11986
- [25] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Cision and Pattern Recognition (CVPR). Las Vegas, USA, 2016: 770-778
- [26] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada, 2021: 10012-10022
- [27] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Miami, USA, 2009: 248-255
- [28] Zhang Z, Zhong S, Liu Y. GANSER: A self-supervised data augmentation framework for EEG-based emotion recognition[J]. *IEEE Transactions on Affective Computing*, early access. doi: 10.1109/TAFFC.2022.3170369
- [29] Chen F, Wang N, Tang J, et al. Self-supervised data augmentation for person re-identification. *Neurocomputing*, 2020, 415: 48-59
- [30] Dong J, Wang W, Tan T. Casia image tampering detection evaluation database//Proceedings of the 2013 IEEE China Summit and International Conference on Signal and Information Processing(ChinaSIP). Beijing, China, 2013: 422-426
- [31] Hsu Y F, Chang S F. Detecting image splicing using geometry invariants and camera characteristics consistency//Proceedings of the 2006 IEEE International Conference on Multimedia and Expo(ICME). Toronto, Canada, 2006: 549-552

- [32] Wen B, Zhu Y, Subramanian R, et al. COVERAGE—A novel database for copy-move forgery detection//Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP). Phoenix, USA, 2016: 161-165
- [33] Guan H, Kozak M, Robertson E, et al. MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation//Proceedings of the 2019 IEEE Winter Applications of Computer Vision Workshops (WACV). Waikoloa Village, USA, 2019: 63-72
- [34] Novozamsky A, Mahdian B, Saic S. IMD2020: a large-scale annotated dataset tailored for detecting manipulated images//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACV). Snowmass Village, USA, 2020: 71-80



**HU Lin-Hui**, M. S. candidate. His current research interests include image forgery detection, deep learning.

**CHEN Bao-Ying**, M. S., engineer. His main research interests include image forensics, video forensics, deep learning.

**TAN Shun-Quan**, Ph. D., associate professor. His main research interests include multimedia security, multimedia forensics, and machine learning.

**LI Bin**, Ph. D., professor. His main research interests include multimedia forensics, image processing, and deep learning.

## Background

Image tampering detection and localization is a crucial research topic in the field of multimedia forensics and an essential means to prevent tampered images from causing misinformation and disrupting social order. With the rapid advancements in deep learning and computer vision, researchers in this area have proposed numerous tampering detection and localization techniques based on deep learning. These methods often demonstrate strong performance on individual datasets, but they tend to possess poor generalization and robustness when applied to different datasets or real-world scenarios.

To address this issue and enhance the model's generalization and robustness, we propose a novel approach by combining Convnext and Upernet, utilizing self-supervised data augmentation and image classification loss functions for image tampering detection and localization. Firstly, Convnext is a state-of-the-art convolutional backbone network that has shown excellent performance in various computer vision tasks. It is based on the ResNet convolutional backbone network and mimics the training approach of the Swin Transformer while outperforming it in terms of performance. By incorporating Convnext into our proposed model, we can efficiently extract and process image features, enabling a better understanding of the underlying image structure. Secondly, Upernet is a well-established model in the scene understanding domain, primarily composed of pyramid pooling modules and feature pyramids. Upernet leverages feature maps of different scales, enhancing the model's detection performance by effectively capturing both

local and global context information. The integration of Upernet allows our model to make better use of multi-scale features, ultimately leading to more accurate tampering detection and localization results. Moreover, we employ self-supervised data augmentation techniques to make tampering traces more evident in images. This process helps the model to learn more diverse and robust representations, which significantly improves its generalization capabilities. Furthermore, we utilize image classification loss functions to optimize the model's performance by minimizing the discrepancy between the predicted tampering probability and the ground truth. A series of ablation experiments have been conducted to demonstrate the effectiveness of the introduced self-supervised data augmentation and image classification loss functions. Our validation experiments on mainstream datasets, such as CASIA, NIST, and COVERAGE, show that the proposed model surpasses existing mainstream models in various evaluation metrics, including AUC, F1 score, and detection accuracy. This model not only performs exceptionally well on individual datasets but also boasts strong generalization and robustness, making it a powerful and reliable tampering detection and localization solution.

This research has been supported by the National Natural Science Foundation of China (U19B2022, 62272314, U22B2047), the Outstanding Youth Project of Guangdong Provincial Natural Science Foundation (2019B151502001), and the Shenzhen Municipal Basic Research Project (JCYJ20200109105008228).