DOI:10.16136/j.joel.2023.04.0340

基于改进的 Mask R-CNN 图像篡改取证方法

吴 云,张玉金*,江潇潇,许灵龙

(上海工程技术大学 电子电气工程学院,上海 201620)

摘要:随着现代科学技术的进步,图像编辑工具的发展极大地降低了篡改所需成本。图像篡改手段有多种,现有的方法往往存在通用性差的问题。同时,这些方法只关注篡改定位而忽略对篡改手段的分类。本文提出一种基于改进的 Mask R-CNN 两阶段网络模型用于图像篡改取证。在特征提取部分,结合空域富模型(spatial rich model,SRM)和受约束卷积对输入图像进行预处理,再输入到 ResNet101 前 4 层中,以建立能够有效体现各种篡改痕迹的统一特征表示。一阶段网络通过注意力区域提议网络(attention region proposal network,A-RPN)检测篡改区域,预测模块实现篡改操作分类和粗略篡改区域定位。继而,一阶段网络得到的定位信息引导二阶段网络学习局部特征以定位出最终的篡改区域。本文所提出的模型能检测 3 种不同类型的图像篡改操作,包括复制-粘贴、拼接和移除。实验结果表明,本文所提出的方法在 NIST16、COVERAGE、Columbia 和 CASIA 数据集的 $F1$ 值分别达到了 0.924、0.761、0.791 和 0.473,优于传统方法和一些主流深度学习方法。

关键词:数字图像取证;篡改检测;深度学习;两阶段网络;注意力机制

中图分类号:TP391 **文献标识码**:A **文章编号**:1005-0086(2023)04-0413-09

A forensic method of image tampering based on improved Mask R-CNN

WU Yun, ZHANG Yujin*, JIANG Xiaoxiao, XU Linglong

(School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

Abstract: With the progress of modern science and technology, the development of image editing tools has greatly reduced the cost of tampering. There are many methods for image tampering, and the existing methods often have the problem of poor universality. Meanwhile, these methods only focus on tampering location and ignore the classification of tampering means. This paper proposes a two-stage network model based on improved Mask R-CNN for image tampering forensics. In the feature extraction part, the input image is preprocessed with spatial rich model (SRM) and constrained convolution, and then input into the first four layers of ResNet101, so as to establish a unified feature representation that can effectively reflect various tampering traces. The first-stage network detects the tampering area through the attention region proposal network (A-RPN), and the prediction module realizes the classification of tampering operation and the location of rough tampering area. Then, the location information obtained by the first-stage network guides the second-stage network to learn local features to locate the final tampering area. The proposed model can detect three different types of image tampering operations, including copy-paste, splicing and removal. The experimental results show that the $F1$ values of the proposed method in NIST16, COVERAGE, Columbia and CASIA datasets reach 0.924, 0.761, 0.791 and 0.473 respectively, which is superior to traditional methods and some state-of-the-art deep learning methods.

Key words: digital image forensics; tampering detection; deep learning; two-stage network; attentional mechanism

* **E-mail**: yjzhang@sues.edu.cn

收稿日期:2022-05-10 **修订日期**:2022-06-19

基金项目:上海市自然科学基金项目(17ZR1411900)和上海市科委重点项目(18511101600)资助项目

0 引　言

数字图像在人们的生活中发挥着至关重要的作用，生活中每个领域都离不开数字图像，例如公共卫生服务、社交媒体平台、司法调查、教育系统、企业等[1]。随着图像/照片编辑工具（如 PhotoShop、美图秀秀等）和其他应用程序的快速发展，人们处理图像和视频逐渐变得方便且快捷。篡改图像的传播影响了人们对客观事件的判断，甚至对社会和国家产生了潜在的威胁。在数字图像篡改手段中，拼接[2,3]、复制-粘贴[4]和移除[5]是为最常见的篡改类型。图像拼接是从一个真实的图像复制区域并粘贴到其他图像上；复制-粘贴是指复制并粘贴区域至同一幅图像中；移除是指移除一个选定的图像区域（例如隐藏一个对象），并用从背景中估计的新像素值来填充空间。图像取证的主要任务是篡改检测和篡改区域定位。篡改检测的目的是识别给定图像是否经过篡改，篡改区域定位是寻找图像中可疑篡改区域。相比而言，后者任务更为复杂。

传统的篡改定位方法主要依赖于手工特征或者预先设定好的特征。BHARTIYA 等[6]提出一种基于特征聚类的方法来解决 JPEG（joint photographic experts group）图像伪造检测的问题。LIU 等[7]采用了一种基于滑动窗口的方法，使用机器学习方法对图像 patch 进行分类来定位篡改区域。BONDI 等[8]提出一种提取相机模型指纹的方法，该方法考虑了特征向量中每一类的置信度分数、纹理程度特征和密度系数。XIONG 等[9]提出一种基于统计噪声分析的方法检测图像拼接，该方法通过估计图像块的噪声值，再通过聚类和两阶段策略定位拼接篡改区域。传统的定位方法需要人工设计特征，且定位准确率低。

近年来，深度学习方法在图像领域中得到了广泛应用。图像取证领域也逐渐从传统的人工特征设计方法转向使用卷积神经网络（convolutional neural network，CNN）自适应地提取图像篡改特征的方法。LIU 等[10]提出一种对抗性学习框架识别拼接篡改区域，该方法使用 3 个构建模块检测拼接的伪造图像。WU 等[11]引入深度匹配和验证网络用于图像拼接篡改取证，该网络能同时实现拼接图像篡改区域的检测和定位。HOSNY 等[12]提出一种轻量级的网络结构用于检测复制-粘贴伪造，该方法在检测精度和效率方面都具有一定优越性。以上方法虽然能取得不错的性能，但是这些方法大多数都集中在单一篡改手段上。

针对以上问题，本文设计了一种适用于多种篡改手段的通用篡改检测网络模型，可以同时实现篡改手段分类和篡改区域定位。该网络首先经过约束卷积和空域富模型（spatial rich model，SRM）结合的图像篡改特征提取器，创建能够有效体现各种篡改手段的统一特征表示，然后再输入到两阶段网络架构中。在两阶段网络中，第一阶段利用注意力区域提议网络（attention region proposal networks，A-RPN）粗略识别篡改区域位置；第二阶段利用跳跃结构融合低层次和高层次特征信息以增强全局特征表示性能，实现最终的篡改区域定位和篡改手段检测。

1　本文方法

1.1　框架概述

本文提出一种通用的、端到端的图像篡改检测网络模型。该网络模型架构如图 1 所示。该网络模型基于 Mask R-CNN 和噪声特征的图像伪造定位方法。为了捕获由图像处理引起的统计特征微小变化，本文将SRM滤波器和约束卷积相结合

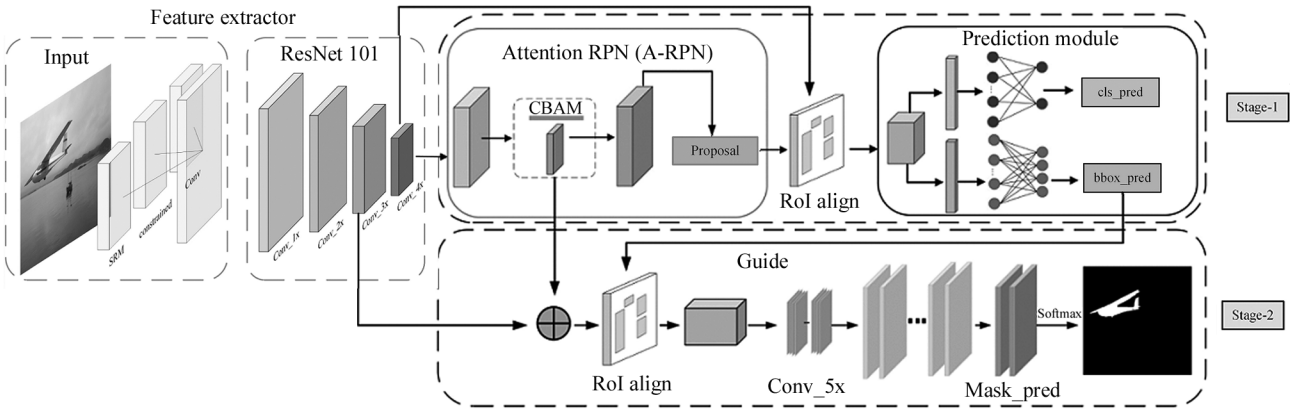

图 1　网络结构示意图

Fig. 1　Schematic diagram of network structure

作为预处理层提取噪声流,ResNet101作为特征提取的骨干网络提取特征。再将提取到的特征输入两阶段网络中,在第一阶段中实现篡改操作分类和粗略篡改区域定位,然后将第一阶段得到的边界框信息引导第二阶段侧重于边界框的局部特征上,实现最终的篡改区域定位。

本文的主要工作体现在以下 3 个方面:

1) 本文将 SRM 高通滤波器与约束卷积层结合作为网络预处理层,利用ResNet101网络提取特征。该方法可以有效地提取噪声特征,提高网络的泛化性能;

2) 本文提出一个基于改进的 Mask R-CNN 两阶段网络模型架构,能同时实现篡改技术检测和篡改区域定位;

3) A-RPN,有效地增强了全局特征的类间区分。在 4 个标准公开数据集上的实验表明,本文的模型优于其他算法。

1.2　网络模型参数

为了更加直观地描述网络结构,在表 1 中详细列出所提出网络模型中每个部分的网络参数。

表 1　网络参数结构表

Tab. 1　Network parameter structure table

Part	Type	Convolution kernels
Feature extractor	SRM	5×5×3
	Constrained Conv	5×5×3
	Common Conv	5×5×10
	Con _1x	7×7,64,stride2
	Max pool	3×3,64,stride2
	Con _2x	[1×1,64;3×3,64;1×1,256]×3
	Con _3x	[1×1,128;3×3,128;1×1,512]×4
	Con _4x	[1×1,256;3×3,256;1×1,1024]×23
Stage-1	Conv	1×1
	A-RPN	
	RoI Align	
	FC	1024
	cls_pred	bbox_pred
Stage-2	RoI Align	
	Conv	1×1
	Conv_5x	[1×1,512;3×3,512;1×1,2 048]×3
	Conv	1×1
	Softmax	

1.3　特征提取

本文将 SRM 滤波器和约束卷积结合提取图像中的噪声残差。卷积网络的第一层起着对输入图像进行预处理的作用。为了验证不同预处理层对图像篡改分类准确性的影响,本文分别对各卷积层进行了对比实验,其结果如表 2 所示。从表 2 中可以看出,将 3 个卷积层串联在一起,网络的检测性能将获得有效的改善。因此,本文选择三者结合作为预处理层,即结合了 3 个 SRM 滤波器、3 个约束卷积滤波器和 10 个普通卷积作为网络预处理层。

表 2　对不同卷积层的比较实验

Tab. 2　Comparative experiments on different convolutional layers

Preprocessing layer	Common Conv	Constrained Conv	SRM	Combined
Number of filters	16	3	3	10+3+3
Kernel Size	(5,5)	(5,5)	(5,5)	(5,5)
Accuracy	0.949	0.938	0.952	0.955

在 30 个基本滤波器中,本文采用文献[13]已知的 SRM 滤波器的最优设置,得到能够有效抑制图像内容的残差。3 个滤波器的权值如图 2 所示。本文将其作为预处理层的卷积之一,核的大小为 5 × 5 × 3。对于彩色图像的输入,通过以上 3 个过滤器,通道数仍为 3。卷积核的权值是固定的,不会随着卷积网络的反向传播而更新。SRM 特征作为手动设计的一个特征,但有其局限性,而约束卷积可以动态地从图像中学习操作特征。具体约束说明如式(1)所示:

$$\begin{cases} w_k^{(1)}(0,0) = -1 \\ \sum_{m,n\neq 0} w_k^{(1)}(m,n) = 1 \end{cases}, \tag{1}$$

式中,w_k 表示第 k 层卷积核,(0,0)表示滤波器矩阵的中心位置。在训练阶段,每次迭代都会遵循约束条件,将(0,0)处的权值设置为−1,除中心外的所有值之和为 1。本文将约束卷积核的大小与 SRM 滤波器设置一样,它的权重自适应地随着网络运行而更新。

$$\frac{1}{4}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 2 & -4 & 2 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \frac{1}{12}\begin{bmatrix} -1 & 2 & -2 & 2 & -1 \\ 2 & -6 & 8 & -6 & 2 \\ -2 & 8 & -12 & 8 & -2 \\ 2 & -6 & 8 & -6 & 2 \\ -1 & 2 & -2 & 2 & -1 \end{bmatrix} \quad \frac{1}{2}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

图 2　3 个 SRM 滤波器的权重设置

Fig. 2　Weight settings of three SRM filters

1.4　一阶段篡改检测

在一阶段篡改检测中，主要任务为篡改手段分类和粗略篡改区域定位。该阶段中包含 A-RPN 模块和预测模块。在预测模块中根据 region proposal 切出感兴趣区域(region of interest，RoI)进行分类和回归预测。

1.4.1　A-RPN

为了增强篡改图像全局特征的类间区分，提升网络模型对篡改图像中篡改区域定位的性能。本文将卷积注意力模块(convolutional block attention module，CBAM)添加到 RPN 中，将特征提取阶段得到的特征图输入到 A-RPN 网络中，使得网络模型能够学习到具有强大的类间区分能力的特征表示。

本文提出的 A-RPN 结构示意图如图 3 所示。A-RPN 利用 CBAM 的特征图来寻找 RoI，即潜在的篡改区域。本文在 1×1 卷积后引入 CBAM，其表达式可以定义为：

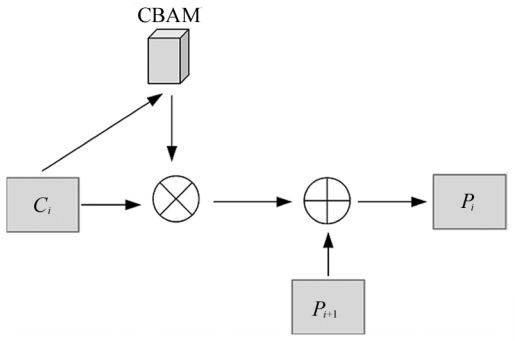

图 3　引入 CBAM 的示意图

Fig. 3　Schematic of introduction of CBAM

$$P_i = C_i \otimes M_i(C_i) \oplus f^{2\times2}_{\text{upsamping}}(P_{i+1}), \tag{2}$$

式中，C_i 表示 ResNet101 特征提取第 i 阶段输出的特征图，P_i 表示对应特征的融合特征图，$\oplus$ 代表特征图对应的元素相加，$\otimes$ 表示特征图对应的元素相乘；$f^{2\times2}_{\text{upsamping}}$ 表示 2×2 的上采样。

为了训练 A-RPN 模块，本文的损失函数是按照文献[14]设计的。A-RPN 的损失函数定义如式(3)所示：

$$L_{\text{A-RPN}}(\{p_i\},\{t_i\}) = \frac{1}{N_{\text{cls}}}\sum_i L_{\text{cls}}(p_i,p_i^*) + \lambda\frac{1}{N_{\text{reg}}}\sum_i p_i^* L_{\text{reg}}(t_i,t_i^*), \tag{3}$$

式中，p_i 表示锚点 i 被判定为对象的概率，p_i^* 是对应于真实标签的概率，t_i 表示预测的边界框，t_i^* 是对应于 ground-truth box 的边界框，L_{cls} 表示分类任务中的损失函数，使用交叉熵分类损失，L_{reg} 表示提议边界框的平滑 L_1 损失，N_{cls} 和 N_{reg} 是 mini-batch 的大小和锚点位置的数量，λ 是一个平衡参数，默认值设置为 10。

1.4.2　预测模块

在 A-RPN 模块之后，通过 RoI Align[15] 裁剪并将 RoI 中的局部特征调整为相同的大小。然后，将全连接层和 Softmax 层用于操作技术分类和边界框回归。这里使用交叉熵损失进行最终操作篡改分类，使用平滑 L_1 损失进行最终边界框回归，实现粗略的篡改定位效果。第一阶段的损失函数定义为：

$$L_{\text{stage-1}} = L_{\text{A-RPN}} + L_{\text{cls\_pred}} + L_{\text{bbox\_pred}}, \tag{4}$$

式中，$L_{\text{A-RPN}}$ 是 A-RPN 的损失，$L_{\text{cls\_pred}}$ 是最终的分类损失，$L_{\text{bbox\_pred}}$ 是最终的边界框回归损失。最后，将 3 个损失相加，得到最终第一阶段的损失函数。

1.5　二阶段篡改检测

在第一阶段中实现了篡改技术分类和粗略篡改区域定位。第二阶段篡改检测主要通过第一阶段的粗略定位信息指导网络模型进一步细化局部特征，并实现篡改图像在像素级别执行篡改区域定位。由于篡改区域边缘存在丰富的伪造信息，本文使用了一个跳跃结构来引入包含更多细节的 Conv_3 中低级特征，并将其与 CBAM 提取的高级特征融合。为了匹配 Conv_5x 块的输入大小，使用核大小为 1×1 的卷积层将全局增强特征的通道扩展到 1024。

同时，第二阶段利用第一阶段的 bounding box 信息进步细化局部特征，然后通过 RoI Align 剪裁局部增强特征并调整大小，再将局部特征输入到 ResNet101 的 Conv_5x 中。最终，利用解码器学习从局部特征到像素预测的映射，实现最终篡改区域定位，也就是利用反卷积层来上采样和减少局部特征的通道。本文使用二元平均交叉熵损失函数计算预测掩码和真实 ground-truth 之间的损失。本文提出模型为端到端的网络模型，输入图像经过特征提取模块以及两阶段网络模块，最终得到预测的篡改内容操作分类和篡改区域二值掩码。最终的损失定义为两阶段的损失之和，定义为 $L_{\text{total}} = L_{\text{stage-1}} + L_{\text{stage-2}}$。其中 L_{total} 为所提出网络模型的损失，$L_{\text{stage-1}}$ 为一阶段所产生的损失，如式(4)所示，$L_{\text{stage-2}}$ 是二阶段平均二元交叉熵损失。

2　实验结果与分析

在本节中，研究所提出模型的有效性以及实现细节，并且列出实验参数。最后在 4 个标准公开数据集上评估所提出模型的性能，并将实验结果与其

他先进方法进行比较。

2.1　数据集

为了评估所提出图像篡改检测方法的性能,本文在 4 个公共篡改数据集上进行实验,包括:NIST16、CASIA、COVERAGE 和 Columbia 数据集。其中 Columbia 数据集不用于训练,仅仅用于测试在 CASIA 数据集上训练的模型。为了增加所提出模型的泛化能力,本文还对数据集进行了数据增强。所提出模型的训练数据集图像数量划分如表 3 所示。

表 3　在 3 个基准数据集中训练和测试图像

Tab. 3　Training and testing images in three benchmark datasets

Datasets	NIST16	CASIA	Columbia	COVERAGE
Training	404	5123	—	75
Testing	160	921	180	25

1) NIST16 数据集包括 564 张篡改图像,经过后处理以隐藏数据集中可见的篡改痕迹。此数据集还提供了 ground-truth 以供评估,它主要包括 3 种篡改类型:复制-粘贴、拼接和移除。该数据集中包含 560 幅原始图像和 564 幅篡改图像,图片格式为 JPG。

2) CASIA 数据集提供各种目标的拼接和复制-粘贴等图像。对被篡改的区域进行了选择并后处理,例如过滤和模糊等。本文使用 CASIA 2.0 (5 123幅图像)进行训练,CASIA 1.0(921 幅图像)用于测试。

3) COVERAGE 数据集是一个相对较小的数据集,专注于图片复制-粘贴。该数据集中包含 100 幅原始图像,每幅图像中都包含两个相近的对象,通过随机复制-粘贴目标对象到同幅图像中,形成复制-粘贴篡改后的图像。

4) Columbia 数据集专注于未压缩图像的拼接。该数据集中包含 363 幅图像,其中有 180 幅拼接篡改图像和 183 幅原始图像,提供了相应的 ground-truth。这些图像包含不同数码相机拍摄的真实图像和拼接后的图像,所有的图片格式均为 TIFF 或 BMP 格式。图像尺寸范围为 757×568—1 152×768 pixel。

2.2　实验参数

由于所提出的网络模型中包含 ResNet101,如果直接将该网络模型进行训练,会出现梯度爆炸现象,所以模型中特征提取和阶段一网络由 ImageNet 权重初始化。本文使用与文献[13]相同的 COCO 合成数据集预训练一阶段网络。预训练网络模型为 110 k 步,最初的学习率设置为 0.001,在 40 k 步后降低为 0.000 1,在 90 k 步后降低为 0.000 01。本文所有的实验是在 NVIDIA GeForce GTX 2080Ti 上进行的,采用 Pytorch 深度学习框架。

2.3　实验结果

2.3.1　评价指标

为了评估所提出模型的有效性,本文使用的性能比较评估指标为像素级 $F1$ 分数和接收器操作特性(receiver operating characteristic,ROC)曲线下的面积(area under curve, AUC),$F1$ 分数是图像处理检测最常用的评价标准。本文使用不同的阈值并使用最高的 $F1$ 分数作为每个图像的最终分数。

$$F1=\frac{2TP}{2TP+FP+FN}, \tag{5}$$

式中,TP(true positive)表示篡改像素被检测正确的数量,FP(false positive)表示实际为真实像素,但是检测结果为篡改像素的数量,TN(true negative)表示真实像素被检测正确的数量,FN(false negative)表示实际为篡改像素,并且检测结果为真实像素的数量。

2.3.2　方法对比

本文评估所提出的网络模型在常见公共数据集上的性能,并且将其与现有的几种传统算法和深度学习方法进行比较,所比较的几种方法如下所示:

1) NOI1[16]方法使用高通小波系数对局部噪声进行建模;

2) ELA[17]方法是分析不同 JPEG 压缩品质并计算其错误级别以确定图片是否有篡改;

3) J-LSTM 方法是基于 LSTM 网络模型,联合训练 patch 篡改边缘分类和像素级篡改区域分割[18];

4) H-LSTM[19]方法使用重采样特征的 LSTM 网络,利用频域特征以及空间内容来定位被篡改的图像区域。

5) RGB-N[13]方法将 RGB 流和通过 SRM 卷积得到的噪声流结合起来,检测图像的篡改区域;

6) MT-Net[20]方法使用自监督学习方法学习篡改特征,并将篡改定位问题作为局部异常值检测问题解决。

本文将所提出模型和以上几种方法进行对比,并列出 $F1$ 分数和 AUC 值。其中 $F1$ 分数的比较如表 4 所示,AUC 值的比较如表 5 所示。

表 4　不同数据集下不同方法的 $F1$ 值比较

Tab. 4　Comparison of $F1$ values of different methods in different datasets

Methods	NIST16	COVERAGE	Columbia	CASIA
NOI1[16]	0.285	0.269	0.574	0.263
ELA[17]	0.236	0.222	0.470	0.214
MFCN[21]	0.570	—	0.612	0.541
RGB-N[13]	0.722	0.437	0.697	0.408
Ours	0.924	0.761	0.791	0.473

表 5　不同数据集下不同方法的 AUC 值比较

Tab. 5　Comparison of AUC values of different methods in different datasets

Methods	NIST16	COVERAGE	Columbia	CASIA
NOI1[16]	0.487	0.587	0.545	0.612
ELA[17]	0.429	0.583	0.581	0.613
J-LSTM[18]	0.764	0.614	—	—
H-LSTM[19]	0.794	0.712	—	—
RGB-N[13]	0.937	0.817	0.858	0.795
MT-Net[20]	0.795	0.819	0.824	0.817
Ours	0.990	0.942	0.863	0.787

表 4 描述了本文方法和其他方法之间 $F1$ 分数的比较，表 5 显示了 AUC 值的比较。从表 4 和表 5 中的结果可以看出，本文所提出的方法在 NIST16 和 COVERAGE 数据集上的检测性能有显著的改善。本文所提出的模型的 $F1$ 分数在 NIST16、COVERAGE 和 Columbia 数据集上分别增长了 0.202、0.324 和 0.094，增长率分别为 28.0%、74.1% 和 13.5%。

在数据增强方面，本文使用了图像翻转、添加高斯噪声和 JPEG 压缩几种方式进行对比。其中高斯噪声的平均值为 0，方差为 5，JPEG 压缩的质量因子为 70。在表 6 中展示了数据增强的实验结果。实验结果表明，图像翻转相比其他数据增强方法能够显著提升准确率。因此本文使用图像翻转作为数据扩充。

表 6　数据增强比较

Tab. 6　Data enhancement comparison

$F1$	COVERAGE	CASIA	NIST16
None	0.748	0.464	0.915
Flipping	0.761	0.473	0.924
Flipping＋Noise	0.743	0.447	0.921
Flipping＋JPEG	0.744	0.484	0.922

2.3.3　篡改手段分类实验结果

图像篡改检测主要包括篡改手段分类和篡改区域定位两个任务，本文模型能够同时完成这两个任务。NIST16 数据集中提供了篡改分类标签：1）复制-粘贴；2）拼接；3）移除。在本文中，使用边界框标识篡改检测分类的结果，如图 5 所示。本文将检测分类结果与文献[13]相比，评价指标为平均精度（average precision，AP），结果如表 7 所示。从表 7 可以看出，本文模型在复制-粘贴上有显著改进，但是在拼接上性能相差不大。

表 7　在 NIST16 数据集上多种类型 AP 比较

Tab. 7　Comparison of multiple types of AP on the NIST16 dataset

AP	Splicing	Copy-move	Removal	Mean
RGB-N[13]	0.960	0.903	0.939	0.934
Ours	0.945	0.996	0.877	0.939

2.3.4　定性分析

在图 4 和图 5 中，本文列出了所提出模型以及 RGB-N[13]、MT-Net[20]的可视化结果。

在图 4 中，由于 COVERAGE、Columbia 和 CASIA 数据集并没有提供篡改类型标签，所以只展示了定位结果的可视化。从图 4 中可以看出，本文模型优于 RGB-N[13]、MT-Net[20]模型，篡改区域和边缘轮廓信息基本都能被检测到，而 RGB-N 方法只能实现块级篡改区域定位，MT-Net 方法定位结果不是很理想。在图 5 中，本文在 NIST16 数据集上展示了多种类型篡改定位结果图，同时在最后一列还列出了篡改操作类型的检测结果。从图 4 和图 5 中可以看出本文所提出的模型有着较好的篡改定位结果和篡改分类检测性能。因为本文将 SRM 滤波器与约束卷积层相结合，有效提取篡改图像噪声特征。同时两阶段网络允许所提出模型在局部区域中执行检测任务，确保获得更准确的篡改区域定位结果。

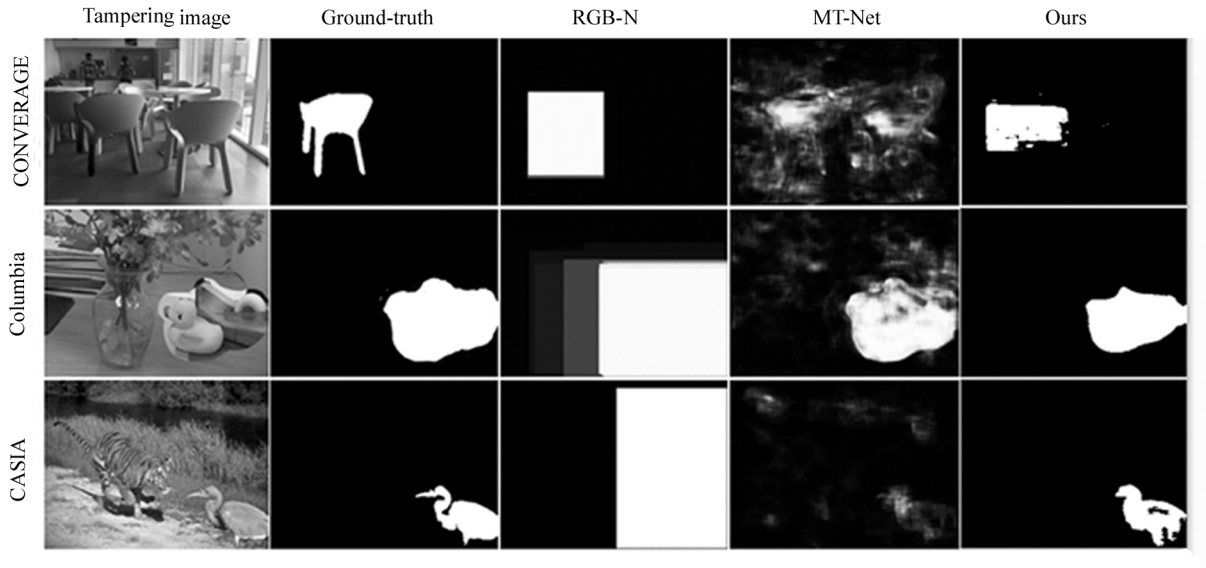

图 4　定位结果的定性可视化

Fig. 4　Qualitative visualisation of localisation results

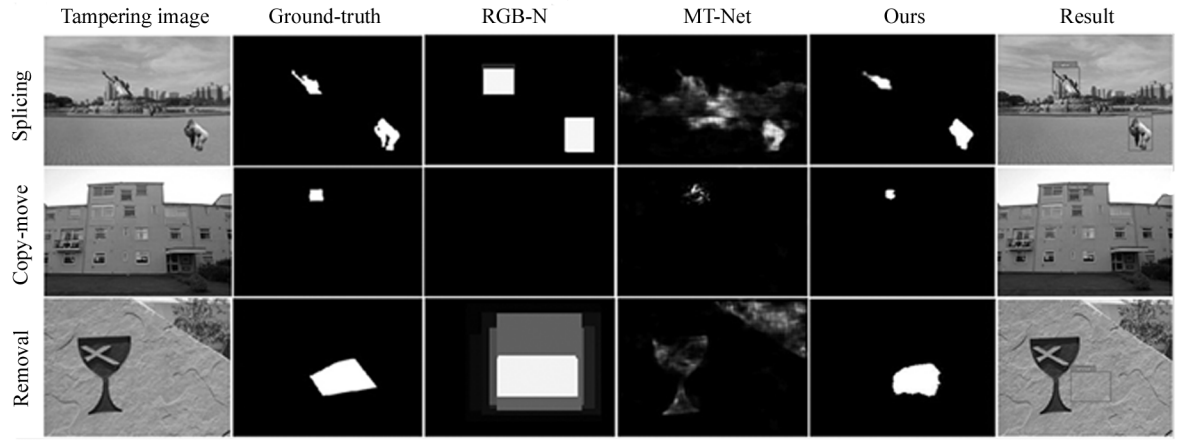

图 5　NIST16 数据集上多种篡改类型的检测结果

Fig. 5　Detection results of multiple tamper types on the NIST16 dataset

2.3.5　消融实验

为了验证本文所提出的预处理层和 A-RPN 模块的有效性,本节对文中的改进点进行消融实验。其中,Base model 表示未添加预处理层和 A-RPN 模块直接用于篡改区域定位;Preprocessing layer 表示使用了 SRM 和受约束卷积相结合作为网络模型的预处理层而没有添加 CBAM 注意力模块;A-RPN 表示添加了 CBAM 注意力模块并没有添加预处理层。在表 8 中列出了 NIST16 和 Columbia 数据集的结果。从表 8 可以看出,同时在网络模型中融入预处理层和 CBAM 模块能有效提高网络对篡改区域定位的精度。

表 8　不同模型检测结果对比(*F*1/AUC)

Tab. 8　Comparison of detection results of different models(*F*1/AUC)

Models	NIST16	Columbia
Base model	0.844/0.901	0.703/0.812
＋Preprocessing layer	0.873/0.950	0.729/0.834
＋A-RPN	0.901/0.963	0.765/0.851
Ours	0.924/0.981	0.791/0.863

2.3.6　复杂度分析

复杂度是衡量一个算法的重要评价指标,也是在实际应用中需要重点考虑的问题之一。本节将所提出的算法与基于深度学习的图像篡改取证方法的

复杂度进行对比分析，结果如表 9 所示。

在表 9 中列出不同方法模型的参数量和平均推理帧率。本文提出的模型参数量仅高于 MT-Net[20] 网络模型，而明显低于 RGB-N[13] 和 MFCN[21] 的参数量。本文提出的算法帧率可达到 13 frame/s，高于其他深度学习方法，所以本文方法能够满足在现实生活中图像篡改检测的实时性需要。

表 9　不同方法复杂度对比

Tab. 9　Complexity comparison of different methods

Methods	Parameters/10^6	Frame/s
RGB-N[13]	1 181.82	0.35
MFCN[21]	512.92	—
MT-Net[20]	72.80	1.25
Ours	382	13

3　结　论

本文提出了一种通用的篡改取证网络模型，该网络模型是基于 Mask R-CNN 网络做出改进，能够同时实现篡改手段分类和篡改区域定位。本文模型将 SRM 和受约束卷积相结合作为预处理层，ResNet101 为骨干网络自适应地学习篡改特征，建立能够有效体现多种篡改手段的统一特征表示，然后通过两阶段网络实现篡改手段分类和篡改区域定位。在 4 个公开数据集上的实验表明，本文所提出的方案的 $F1$ 分数和 AUC 值均取得较好的结果，优于传统方法以及大多数深度学习方法，并且在多种篡改数据集上有着较好的泛化能力。在未来的工作中，将探索篡改取证中的小目标和多目标检测，提升网络模型对复杂图像取证的通用性和有效性。

参考文献：

[1] LIN X，LI J H，WANG S L，et al. Recent advances in passive digital image security forensics：A brief review[J]. Engineering，2018，4(1)：29-39.

[2] BI X，WEI Y，XIAO B，et al. RRU-Net：The ringed residual U-Net for image splicing forgery detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops，June 16-17，2019，Long Beach，CA，USA. New York：IEEE，2019：30-39.

[3] Chen B，Qi X，Wang Y，et al. An improved splicing localization method by fully convolutional networks[J]. IEEE Access，2018，6：69472-69480.

[4] WEI W Y，WANG L Z，WANG W R，et al. Image copy-move forgery detection algorithm based on superpixel shape features[J]. Computer Engineering and Science，2021，43(2)：312-321.
魏伟一，王立召，王婉茹，等. 基于超像素形状特征的图像复制粘贴篡改检测算法[J]. 计算机工程与科学，2021，43(2)：312-321.

[5] ZHU X，QIAN Y，ZHAO X，et al. A deep learning approach to patch-based image inpainting forensics[J]. Signal Processing：Image Communication，2018，67：90-99.

[6] BHARTIYA G，JALAL A S. Forgery detection using feature-clustering in recompressed JPEG images[J]. Multimedia Tools and Applications，2017，76(20)：20799-20814.

[7] LIU Y，GUAN Q，ZHAO X，et al. Image forgery localization based on multi-scale convolutional neural networks[C]//Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security，June 20-22，2018，New York，United States. New York：ACM，2018：85-90.

[8] BONDI L，LAMERI S，GUERA D，et al. Tampering detection and localization through clustering of camera-based CNN features[C]//CVPR Workshops，July 21-26，2017，Honolulu，HI，USA. New York：IEEE，2017：1855-1864.

[9] XIONG S T，ZHANG Y J，WU F，et al. Image splicing detection based on statistical noise level analysis[J]. Journal of Optoelectronics · Laser，2020，31(2)：214-221.
熊士婷，张玉金，吴飞，等. 基于统计噪声水平分析的图像拼接检测[J]. 光电子 · 激光，2020，31(2)：214-221.

[10] LIU Y，ZHU X，ZHAO X，et al. Adversarial learning for constrained image splicing detection and localization based on atrous convolution[J]. IEEE Transactions on Information Forensics and Security，2019，14(10)：2551-2566.

[11] WU Y，ABD-ALMAGEED W，NATARAJAN P. Deep matching and validation network：An end-to-end solution to constrained image splicing localization and detection[C]//Proceedings of the 25th ACM International Conference on Multimedia，May 27，2017，Shanghai，China. New York：ACM，2017：1480-1502.

[12] HOSNY K M，MORTDA A M，FOUDA M M，et al. An efficient CNN model to detect copy-move image forgery[J]. IEEE Access，2022，10：48622-48632.

[13] ZHOU P，HAN X，MORARIU V I，et al. Learning rich features for image manipulation detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，June 18-23，2018，Salt Lake City，UT，USA. New York：IEEE，2018：1053-1061.

[14] REN S，HE K，GIRSHICK R，et al. Faster R-CNN：towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2017，39(6)：1137-1149.

[15] HE K，GKIOXARI G，DOLLáR P，et al. Mask R-CNN[C]//

Proceedings of the IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 2961-2969.

[16] MAHDIAN B, SAIC S. Using noise inconsistencies for blind image forensics[J]. Image and Vision Computing, 2009, 27(10): 1497-1503.

[17] KRAWETZ N, SOLUTIONS H F. A pictures worth[J]. Hacker Factor Solutions, 2007, 6(2): 2.

[18] BAPPY J H, ROY-CHOWDHURY A K, BUNK J, et al. Exploiting spatial structure for localizing manipulated image regions[C]//Proceedings of the IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 4970-4979.

[19] BAPPY J H, SIMONS C, NATARAJ L, et al. Hybrid ISTM and encoder-decoder architecture for detection of image forgeries[J]. IEEE Transactions on Image Processing, 2019, 28(7): 3286-3300.

[20] WU Y, ABDALMAGEED W, NATARAJAN P. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 15-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 9543-9552.

[21] SALLOUM R, REN Y Z, KUOC C J, et al. Image splicing localization using a multi-task fully convolutional network (MFCN)[J]. Journal of Visual Communication & Image Representation, 2018, 51: 201-209.

作者简介:

张玉金　(1982),男,博士,副教授,硕士生导师,主要从事多媒体取证、信号处理、人工智能和模式识别方面研究.