Computer Science and Application 计算机科学与应用, 2022, 12(3), 729-738
Published Online March 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.123074

Hans汉斯

基于注意力机制的图像篡改检测

余　晨[1]，符　颖[1,2]，朱　烨[1]

[1]成都信息工程大学计算机学院，四川　成都
[2]四川省图形图像与空间信息2011协同创新中心，四川　成都

收稿日期：2022年2月24日；录用日期：2022年3月22日；发布日期：2022年3月30日

摘　要

目前已有的基于分割的图像篡改检测方法由于标注困难，可用的篡改数据集较少，造成训练数据的缺乏，同时篡改图像经过处理后边界难以识别，导致分割精度低。针对上述问题提出了基于注意力机制的图像篡改检测网络，该网络实现了篡改图像的生成，篡改区域的分割和优化。其中，生成器创建篡改图像用于扩充训练数据，基于注意力机制的分割优化模块用于增强篡改区域边界的特征提取能力，最后以实验结果证明了此方法的准确性和有效性。

关键词

图像篡改，分割，注意力机制

Image Tampering Detection Based on Attention Mechanism

Chen Yu[1], Ying Fu[1,2], Ye Zhu[1]

[1]School of Computer Science, Chengdu University of Information Technology, Chengdu Sichuan
[2]2011 Collaborative Innovation Center of Graphics, Image and Spatial Information of Sichuan Province, Chengdu Sichuan

Received: Feb. 24th, 2022; accepted: Mar. 22nd, 2022; published: Mar. 30th, 2022

Abstract

At present, the existing image tamper detection methods based on segmentation are difficult to label and have few available tamper data sets, resulting in the lack of training data. At the same time, the boundary of the tampered image is difficult to identify after processing, resulting in low segmentation accuracy. To solve the above problems, an image tamper detection network based on attention mechanism is proposed. The network realizes the generation of tampered images, the

文章引用: 余晨, 符颖, 朱烨. 基于注意力机制的图像篡改检测[J]. 计算机科学与应用, 2022, 12(3): 729-738.
DOI: 10.12677/csa.2022.123074

segmentation and optimization of tampered regions. The generator creates the tampered image to expand the training data, and the segmentation optimization module based on attention mechanism is used to enhance the feature extraction ability of the tampered region boundary. Finally, the experimental results show the accuracy and effectiveness of this method.

Keywords

Image Tampering, Segmentation, Attention Mechanism

1. 引言

由于图像编辑软件的简单实用性，人们可以随意对图像进行篡改，导致篡改图像在社交媒体上无处不在。人们篡改图像的目的多种多样，有些是为了美化形象、丰富图像内容，而有些则是为了散布虚假消息引起不良的社会影响。为了防止此类虚假消息的广泛传播，图像篡改检测也随之兴起，数字图像篡改的取证技术主要分为主动取证技术和被动取证技术。主动取证方法主要包含数字签名[1]、数字水印[2]等需要事先在图像中嵌入验证信息，然后再进行信息提取验证的相关技术手段。相对于主动取证，被动取证的应用范围更为广泛，可以直接通过图像进行图像来源、完整性的判断。

关于传统方法的图像被动取证，针对 JPEG 压缩操作，Fridrich 等人[3]首先对图像进行分块处理，再以图像块为单位提取特征。H. Farid 等人[4]提出主成分分析法(PCA)代替离散余弦变换特征系数(DCT)，获得更小的特征系数，提升了图像块特征向量匹配的效率。传统的取证方法通常都是采用手工设计的方式提取特征，这种基于手工设计的特征大多存在局限性，缺乏代表性，无法根据这些特征同时对多种篡改方式进行判定，导致传统的被动取证方法仅能鉴别单种篡改技术。

关于深度学习方法的图像被动取证，Yaqi Liu 等人[5]提出了一种基于多尺度方法进行篡改区域定位的框架。该模型创新性地引入了多尺度的思想，设计不同尺度的图像滑动窗口，提取多尺度图像块特征。Bo Liu 等人[6]提出了一种多种基础网络融合的方法，利用不同种类的基础网络互补，用来提升模型的鲁棒性，从而提升模型的识别准确率。Minyoung Huh 等人[7]创新性地将自学习方法融入图像拼接检测中，该模型避免了当前图像篡改数据集数据量不足的问题，为基于深度学习的图像内容被动取证指出了新的方向。通过研究发现，基于深度学习的图像内容被动取证领域还存在以下问题：1) 缺乏数据集，虽然当前篡改图像数据集越来越多，但这与复杂多样的篡改技术相比仍远远不足。2) 提取篡改区域特征困难，在数字图像处理中，根据不同目的对图像进行不同处理，例如图像去噪、增强、复原和压缩等，而这些后处理技术使得提取图像篡改特征更加困难。

基于上述问题，本文提出一种新的图像篡改检测分割框架，框架分为两个阶段：第一阶段采用 GAN [8]来解决缺乏训练数据的问题。通过引入一个新的目标函数，将来自现有篡改数据集的图像作为 GAN 的输入，并通过目标函数进行优化，再与原篡改图像数据集进行混合，从而达到扩充数据集与优化篡改区域真实性的效果。第二阶段以 Deeplab [9]为基本框架，提出基于注意力机制的分割优化网络，将卷积注意力模块[10] (CBAM)加入分割网络中，以增强篡改区域边界的特征提取能力，得到边缘 mask 和篡改区域 mask，同时利用预测的边缘 mask 来替换原始区域，产生新的篡改图像，强化模型对于边界特征的提取能力，从而更好地定位篡改区域并将其分割出来。本文的主要贡献如下：

1) 针对缺乏训练数据的问题，通过 GAN 来扩充数据集并引入新的目标函数用于优化篡改区域真实性。

2) 针对篡改区域边界识别困难以及分割准确率低的问题，提出基于注意力机制的分割优化网络，提高模型性能。

2. 方法

本文所提出的图像篡改检测网络结构如图 1 所示，第一阶段将篡改数据集的图像和真实图像经过处理后作为生成器的输入，并通过目标函数进行优化，生成更加真实的篡改图像，并送入判别器进行验证；第二阶段再将生成的图像与原篡改图像数据集进行混合，作为注意力分割模块的输入，输出预测的边缘 mask 和篡改区域 mask。其中，生成器基于 U-net [11]网络结构，判别器与 PatchGAN [12]类似；注意力分割模块分为分割模块和优化模块，分割模块负责识别篡改图像中的篡改边界，优化模块将篡改边界替换为真实区域，生成新的篡改边界。

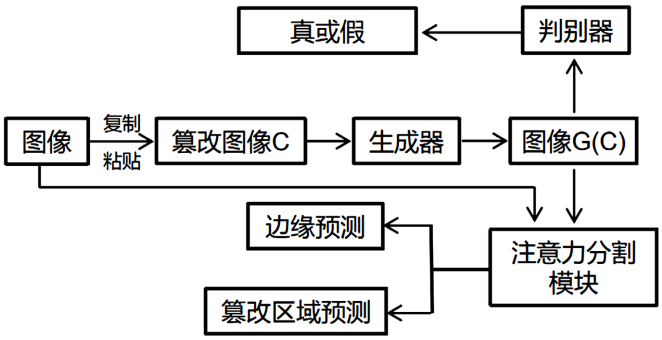

Figure 1. The network structure
图 1. 网络结构图

2.1. 优化和扩充数据集

2.1.1. 生成器

生成器生成的篡改图像作为输入用于后续分割训练，如图 2 所示。根据现有数据集提供的篡改图像 F (CASIA2.0 [13])以及对应的 mask 图像 M 和真实图像 R (COCO [14])，首先以图像 F 为前景，图像 R 为背景创建一个复制粘贴图像 C，C 的表示如公式(1)所示：

$$C=(1-M)*R+M*F \tag{1}$$

其中 $*$ 是按元素进行的矩阵乘法。

混合图像通过泊松融合[15]得到篡改区域中像素 i 的最终值如公式(2)所示：

$$a_i=\arg_{a_i}\min\sum_{f_i\in F,N_i\subset F}\left\|\nabla a_i-\nabla f\right\|_2+\sum_{f_i\in F,N_i\not\subset F}\left\|a_i-r_i\right\|_2 \tag{2}$$

其中 ∇ 表示梯度，N_i 是位置 i 处的像素的邻域，a_i 是混合图像 A 中的像素，f_i 是图像 F 中的像素，r_i 是图像 R 中的像素。

与泊松融合类似，图像 C 中包含复制粘贴区域和背景图像，为了更好的融合篡改区域的邻域，本文优化了生成器，并添加了 L_1 损失[12]来重建背景，背景损失如公式(3)所示：

$$L_{bg}=\frac{1}{N_{bg}}\sum_{r_i\in R,m_i=0}\left\|c_i-r_i\right\|_1 \tag{3}$$

其中 N_{bg} 是背景中的像素总数，c_i 是图像 C 中的像素，m_i 是图像 M 中位置 i 处的值。

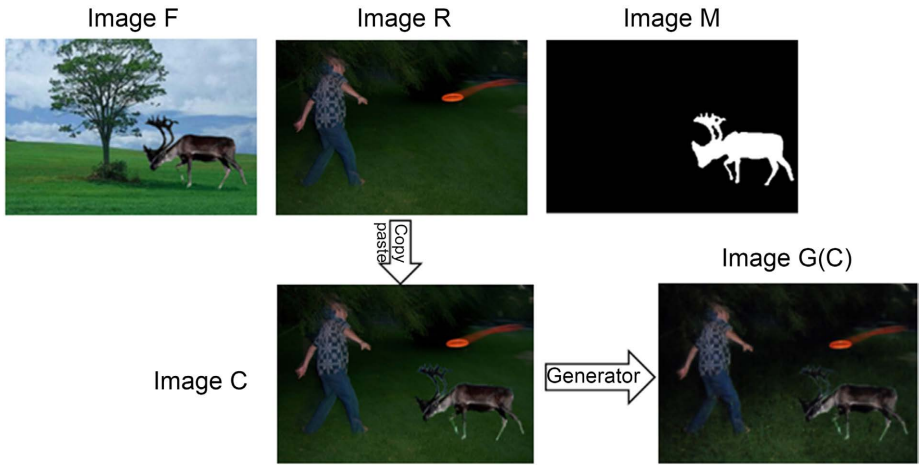

Figure 2. Generate image
图 2. 生成图像

为了维持篡改区域的形状，本文对复制粘贴的区域应用拉普拉斯算子，并重建该区域的梯度以匹配原区域，梯度损失如公式(4)所示：

$$L_{grad} = \frac{1}{N_{copy}} \sum_{f_i \in F, m_i = 1} \left\| \Delta c_i - \Delta f_i \right\|_1 \tag{4}$$

其中，Δ 表示拉普拉斯运算子，N_{copy} 是复制粘贴区域中的像素总数。

为了进一步约束复制粘贴区域的形状，本文添加的边缘损失如公式(5)所示：

$$L_{edge} = \frac{1}{N_{edge}} \sum_{f_i \in F, e_i = 1} \left\| c_i - f_i \right\|_1 \tag{5}$$

其中 N_{edge} 是边界像素的数量，e_i 是边缘 mask 中位置 i 处的值，边缘 mask 是通过对图像 M 进行膨胀和腐蚀操作得到的。

2.1.2. 判别器

由于篡改图像中被篡改的区域通常只占据了图像的一小部分区域，因此，限制判别器注意局部图像块中的特征是有益的。与 PatchGAN [12]类似，本文的判别器在 $N \times N$ 的图像块上应用全卷积层。判别器通过最大化损失判断真实图像 R 为真，判断生成的图像 $G(M,C)$ 为假。为了生成真实的篡改图像，本文添加了一个对抗损失 L_{adv}，这有助于鼓励生成器在训练的过程中生成越来越真实的图像，对抗损失如公式(6)所示：

$$L_{adv} = E_R\left[\log\left(D(M,R)\right)\right] + E_C\left[1 - \log\left(D\left(M,G(M,C)\right)\right)\right] \tag{6}$$

其中，M 与 $G(M,C)$ 或 R 相连，作为判别器的输入。生成器最终损失函数如公式(7)所示：

$$L_G = L_{bg} + \lambda_{grad} L_{grad} + \lambda_{edge} L_{edge} + \lambda_{adv} L_{adv} \tag{7}$$

其中，λ_{grad}、λ_{edge} 和 λ_{adv} 分别设置为(1, 2, 5)以控制相应损失项。基于此约束，生成器保留并融合复制粘贴区域的背景以及纹理信息。

2.2. 注意力分割模块

2.2.1. 分割模块

分割网络结构如图 3 所示，分割阶段分成两个分支，一是预测篡改区域边缘 mask 的边缘分支，二是预测篡改区域 mask 的分割分支。其中为了提高对边界的关注，在分割网络中加入卷积注意力模块 CBAM，

CBAM 通过空间和通道两个维度依次推断出注意力权重，然后与特征图相乘来对特征进行自适应调整，再将注意力图输入边缘分支和分割分支。随后将双线性上采样的中间特征连接起来，再将它们输入到 1 × 1 的卷积层从而输出预测的边缘 mask，以形成边界分支。最后，本文将边界分支的输出特征与分割分支的上采样特征结合起来，输出最终的篡改区域 mask。在分割网络的训练过程中，本文选择了复制粘贴的样本 C，生成的样本 $G(C)$和训练样本 F 作为分割网络的输入，提供了更多种类的篡改图像。

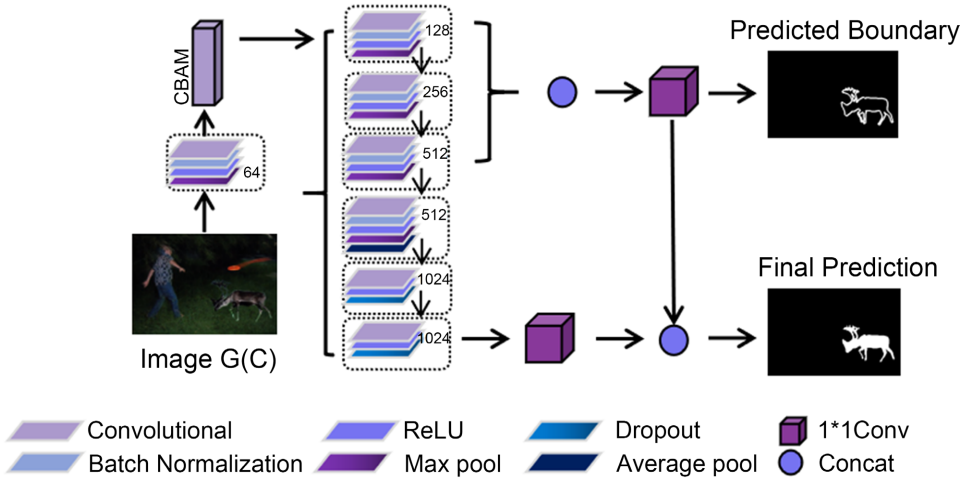

Figure 3. Generate image
图 3. 生成图像

2.2.2. 优化模块

语义图像分割注重的是图像的语义内容，而篡改区域的分割更注重于篡改边界，如图 4 所示，本文利用分割出来的边缘 mask，替换原始篡改区域的边界，生成新的篡改边界。本文用 R 中的真实区域替换预测边界中的像素，生成一个新的篡改图像，C' 的表示如公式(8)所示：

$$C' = C \cdot (1 - P) + R \cdot P \tag{8}$$

其中，C' 是具有新边界伪影的新篡改图像。对应的 mask 如公式(9)所示：

$$M' = M - M \cdot P \tag{9}$$

其中，M' 是 C' 的篡改区域 mask。用上述相同的方式预测新的边缘 mask 和篡改区域 mask。

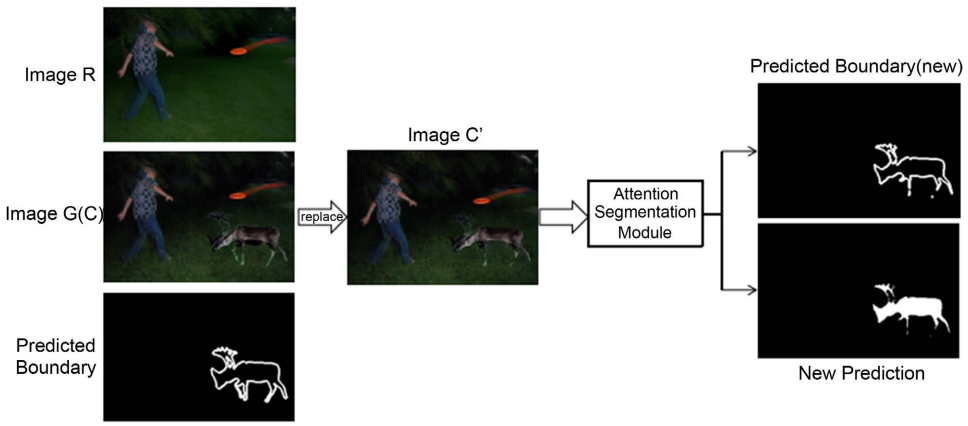

Figure 4. Optimization module
图 4. 优化模块

3. 实验

3.1. 实验环境

实验环境如表 1 所示。由于模型参数较大，因此对实验的设备有一定要求，本实验采用部署了英伟达 RTX 2070SUPERx 显卡的 Linux 服务器进行实验，并且配置了符合要求的加速平台和加速库。初始化学习率设置为 0.0001，batch size 设置为 2，网络优化方法使用 Adam，动量为 0.9，权重衰减设置为 0.0005。

Table 1. Experimental environment
表 1. 实验环境

设备	配置	备注
操作系统	Ubuntu	16.04
框架	Tensorflow	1.4.0
GPU	Nvidia RTX 2070SUPER	8GB
运行平台	CUDA	8.0.44
GPU 加速库	Cudnn	6.0
语言	Python3	3.4

3.2. 实验对比

3.2.1. 数据集与评价指标

本文提出的方法在 CASIA 数据集和 Columnbia 数据集上与其他经典的网络进行了比较。CASIA1.0 包含 921 个篡改图像，CASIA2.0 包含 5123 张图像，篡改技术包括拼接和复制粘贴，被篡改的区域包括动物、纹理、自然场景等，并且该数据集对篡改图像进行了后处理，以增强篡改效果。Columnbia 数据集包含 183 张拼接图像，被篡改的区域是室内场景。本文使用 CASIA2.0 对模型进行训练，在 CASIA1.0 和 Columnbia 上进行测试。本文使用指标 F_1 分数(F1-Score)和 MCC 分数(MCC-Score)评估本文提出的方法和现有方法的性能，它们都可以用来评价篡改区域的检测精度。F1-Score 是查全率与查准率的加权调和平均，MCC-Score 则是一个比较均衡的指标，在两类别的样本含量差别很大的情况下可以更好的对模型进行评价，比较适用于篡改检测领域。评估指标 F1-Score 如公式(10)所示：

$$F_1\left(M_{out}, M_{gt}\right) = \frac{2TP}{2TP + FN + FP} \tag{10}$$

其中 M_{out} 代表算法输出 mask，M_{gt} 代表真实 mask。TP 代表篡改区域检测为篡改像素数量。FN 代表未篡改区域检测为篡改像素数量。FP 代表篡改区域未检测为篡改像素数量。评估指标 MCC-Score 如公式(11)所示：

$$MCC\left(M_{out}, M_{gt}\right) = \frac{TP * TN - FP * FN}{\sqrt{\left(TP + FP\right)\left(TP + FN\right)\left(TN + FP\right)\left(TN + FN\right)}} \tag{11}$$

3.2.2. 实验结果

从表 2 和表 3 可以看出，本文所提出的方法在 F1-Score 和 MCC-Score 方面优于现有的一些方法。如判断图像来源的被动取证方法 NOI [16]和 CFA [17]，这是因为手工设计的特征限制了它们的功能。本文引入的 CBAM 模块在不同空间生成上下文描述符，有助于获得篡改图像中的边界特征。

Table 2. Results on F1-Score
表 2. 实验结果对比(F1-Score)

	CASIA	Columbia
NOI	0.263	0.574
CFA	0.207	0.467
MFCN [18]	0.518	0.604
GSRnet [19]	0.574	0.532
Our method	**0.602**	**0.623**

Table 3. Results on MCC-Score
表 3. 实验结果对比(MCC-Score)

	CASIA	Columbia
NOI	0.180	0.411
CFA	0.108	0.228
MFCN [18]	0.484	0.465
GSRnet [19]	0.553	0.496
Our method	**0.581**	**0.537**

从图 5 可以看出本文提出的方法能够有效的将篡改区域分割出来，分割的边缘 mask 也比较清晰，虽然篡改区域经过后处理，隐藏了篡改区域和真实区域的对比差异，但本文通过在分割网络中添加 CBAM 模块，增强了篡改区域边界的特征提取能力，使得拼接类型的篡改图像更容易检测，分割效果也较好。而对于复制粘贴类的篡改图像检测则较困难，这主要是因为复制区域来自于同一张篡改图像的某个部分，篡改区域与真实区域有着相同的对比度，在本文设计的网络中，通过分割模块中的优化部分，强化模型对于边界特征的提取能力，使得复制粘贴类型的篡改图像也容易被检测与分割，本文提出的方法在两种篡改图像上都取得了较好的效果。

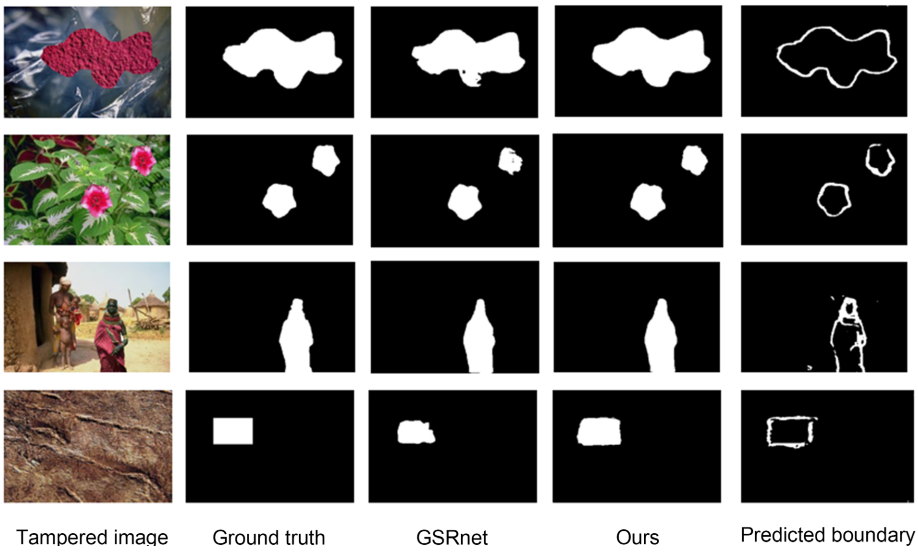

Figure 5. The segmentation result
图 5. 分割结果图

3.2.3. 鲁棒性分析

为验证模型的鲁棒性，对 CASIA 数据集采用 JPEG 压缩和高斯平滑处理，其中，JPEG 压缩将图片质量压缩到 70%和 50%，高斯平滑处理将高斯核标准差设置为 0.5、1、1.5 和 2。从图 6 和图 7 可以看出，模型的抗干扰性较强，有较强的鲁棒性。

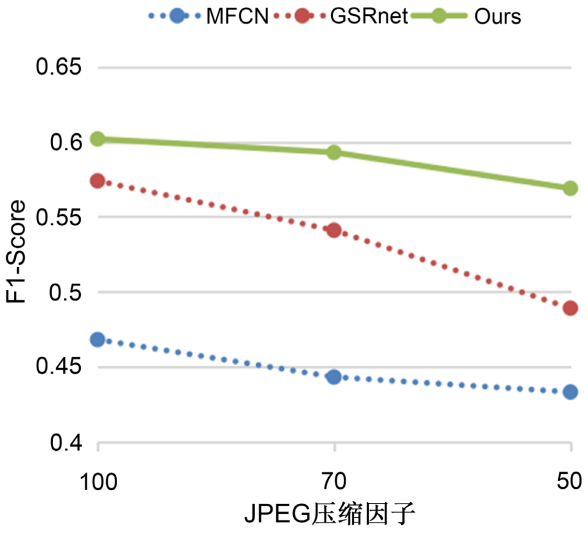

Figure 6. JPEG compression
图 6. JPEG 压缩

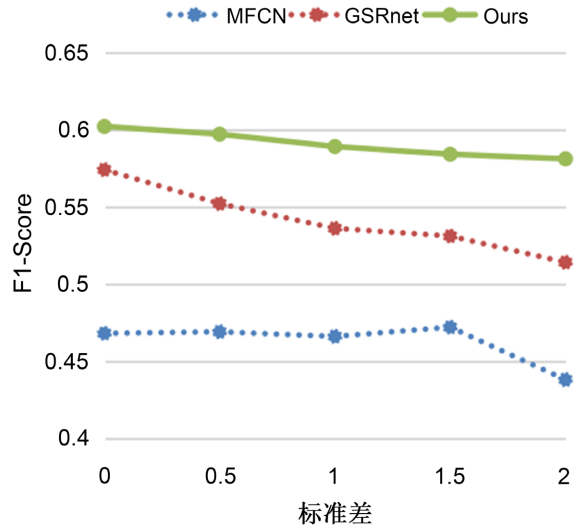

Figure 7. Gaussian smoothing
图 7. 高斯平滑处理

4. 结束语

本文提出了一种基于注意力机制的图像篡改检测网络，实现了图像篡改区域的分割。利用 GAN 生成篡改图像来解决缺乏训练数据的问题，同时扩充样本量避免出现模型坍塌的情况。相较于目前的生成对抗网络，本文引入了一个新的目标函数，将来自现有篡改数据集的图像作为 GAN 的输入，并通过目标函数进行优化，再与原篡改图像数据集进行混合，从而达到扩充数据集与优化篡改区域真实性的效果。为

了在训练过程中使模型更好地识别边界，进一步提取有效的图像篡改特征，本文以 Deeplab 为基本框架，提出了基于注意力机制的分割优化网络，将卷积注意力模块(CBAM)加入分割网络中，以增强篡改区域边界的特征提取能力，得到边缘 mask 和篡改区域 mask，同时利用预测的边缘 mask 来替换原始区域以产生新的篡改图像，强化模型对于边界特征的提取能力，从而更好地定位篡改区域并将其分割出来。从对比实验可以看出，本文方法各项指标均优于其它方法。

由于网络模型参数较多，对计算机的运算能力要求较高，后期的工作主要在于构建轻量化的模型并提高算法的检测精度。

基金项目

本文作者符颖得到省级项目基金资助：四川省科技厅重点研发项目(2020YFG0453，8K 超短焦光学镜头关键技术研究及应用)；四川省科技厅“新一代人工智能平台”重大专项(2019DZX0005，面向开放共享的云深度学习专用平台)。

参考文献

[1] Lu, C.S., *et al*. (2003) Structural Digital Signature for Image Authentication: An Incidental Distortion Resistant Scheme. *IEEE Transactions on Multimedia*, **5**, 161-173. https://doi.org/10.1109/TMM.2003.811621

[2] Weng, S., Yao, Z., Pan, J.S., *et al*. (2008) Reversible Watermarking Based on Invariability and Adjustment on Pixel Pairs. *IEEE Signal Processing Letters*, **15**, 721-724. https://doi.org/10.1109/LSP.2008.2001984

[3] Fridrich, A.J., Soukal, B.D. and Luk, S.A.J. (2003) Detection of Copy-Move Forgery in Digital Images. *Proceedings of Digital Forensic Research Workshop*, Cleveland, 6-8 August 2003, 55-61.

[4] Popescu, A.C. and Farid, H. (2004) Exposing Digital Forgeries by Detecting Duplicated Image Regions. Department of Computer Science, Dartmouth College, Hanover, Tech. Rep. TR2004-515, 1-11.

[5] Liu, Y., Guan, Q., Zhao, X., *et al*. (2018) Image Forgery Localization Based on Multi-Scale Convolutional Neural Networks. *Proceedings of the* 6*th ACM Workshop on Information Hiding and Multimedia Security*, Innsbruck, 20-22 June 2018, 85-90. https://doi.org/10.1145/3206004.3206010

[6] Liu, B. and Pun, C.-M. (2018) Deep Fusion Network for Splicing Forgery Localization. *Proceedings of the European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 237-251. https://doi.org/10.1007/978-3-030-11012-3_21

[7] Huh, M., Liu, A., Owens, A., *et al*. (2018) Fighting Fake News: Image Splice Detection via Learned Self-Consistency. *Proceedings of the European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 106-124. https://doi.org/10.1007/978-3-030-01252-6_7

[8] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) Generative Adversarial Nets. *Advances in Neural Information Processing Systems* 27: *Annual Conference on Neural Information Processing Systems* 2014, Montreal, 8-13 December 2014, 2672-2680.

[9] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A.L. (2018) DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**, 834-848. https://doi.org/10.1109/TPAMI.2017.2699184

[10] Woo, S., Park, J., Lee, J.Y., *et al*. (2018) Cbam: Convolutiona Block Attention Module. *Proceedings of the European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[11] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-net: Convolutional Networks for Biomedical Image Segmentation. 18*th International Conference on Medical Image Computing and Computer-Assisted Intervention*, Munich, 5-9 October 2015, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[12] Isola, P., Zhu, J.-Y., Zhou, T. and Efros, A.A. (2017) Image-to-Image Translation with Conditional Adversarial Networks. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 5967-5976. https://doi.org/10.1109/CVPR.2017.632

[13] Dong, J., Wang, W. and Tan, T. (2013) Casia Image Tampering Detection Evaluation Database. 2013 *IEEE China Summit and International Conference on Signal and Information Processing*, *ChinaSIP* 2013, Beijing, 6-10 July 2013, 422-426. https://doi.org/10.1109/ChinaSIP.2013.6625374

[14] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P. and Zitnick, C.L. (2014) Microsoft Coco: Common Objects in Context. *Computer Vision—ECCV* 2014: 13*th European Conference*, Zurich, 6-12 September 2014, 740-755. https://doi.org/10.1007/978-3-319-10602-1_48

[15] Perez, P., Gangnet, M. and Blake, A. (2003) Poisson Image Editing. *ACM Transactions on Graphics*, **22**, 313-318. https://doi.org/10.1145/1201775.882269

[16] Mahdian, B. and Saic, S. (2009) Using Noise Inconsistencies for Blind Image Forensics. *Image and Vision Computing*, **27**, 1497-1503. https://doi.org/10.1016/j.imavis.2009.02.001

[17] Ferrara, P., Bianchi, T., De Rosa, A. and Piva, A. (2012) Image Forgery Localization via Fine-Grained Analysis of CFA Artifacts. *IEEE Transactions on Information Forensics and Security*, **7**, 1604-1613. https://doi.org/10.1109/TIFS.2012.2202227

[18] Salloum, R., Ren, Y. and Kuo, C. (2017) Image Splicing Localization Using a Multi-Task Fully Convolutional Network (MFCN). *Journal of Visual Communication & Image Representation*, **51**, 201-209. https://doi.org/10.1016/j.jvcir.2018.01.010

[19] Zhou, P., Chen, B.-C., Han, X., Najibi, M., Shrivastava, A., Lim, S.-N. and Davis, L. (2020) Generate, Segment, and Refine: Towards Generic Manipulation Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 13058-13065. https://doi.org/10.1609/aaai.v34i07.7007