第 58 卷 第 22 期/2021 年 11 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于改进 RGB-N 的图像操纵检测算法

刘昊岳*，马文伟，付晓，沈程秀，王亚领

泰康在线财产保险股份有限公司互联网金融实验室，湖北 武汉 430014

摘要 如何准确地检测出图像中的操纵痕迹是数字图像被动取证领域的研究重点。传统方法利用人工构造的特征进行检测，鲁棒性不强，而基于深度学习的方法虽具有较强的检测能力，但较少关注在正常图像上出现误检的情况。提出了一种改进的 RGB-N 图像操纵检测算法，该算法在使用 F1 分数评价操纵目标检测性能的同时，引入了在正常图像上的误检率指标来评价算法的实用性。设计了自适应空域富模型滤波器，构造多尺度融合的特征提取网络，并接入自注意力模块，增强了模型获取图像全局信息的能力，提高检测性能；为降低误检率，设计了真实性判断模块，输出的热图用于判断检测到的目标是否为误检，并通过从操纵目标来源图像选择负样本的训练策略进一步提高模型的分辨能力。实验结果表明，改进的 RGB-N 模型在含目标拼接与擦除两种操纵手段的数据集上的 F1 分数为 0.759，在未操纵图像数据集上的误检率为 0.2%，并在 JPEG 压缩攻击下具有较好的鲁棒性。

关键词 成像系统；图像操纵检测；自适应空域富模型；多尺度融合；自注意力；真实性判断

中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP202158.2211003

Image Manipulation Detection Algorithm Based on Improved RGB-N

Liu Haoyue*, Ma Wenwei, Fu Xiao, Shen Chengxiu, Wang Yaling

Internet Finance Laboratory, TK.CN Insurance Co., Ltd., Wuhan, Hubei 430014, *China*

Abstract How to accurately detect the manipulation trace in images is the research focus in digital image passive forensics. Traditional methods use artificial features to detect, without enough robustness. Although the method based on deep learning has strong detection ability, it pays less attention to false detection on normal images. An improved RGB-N image manipulation detection algorithm is proposed. The algorithm uses F1 score to evaluate the detection performance of manipulation targets, and introduces the false detection rate index on normal images to evaluate the practicability of the algorithm. An adaptive spatial rich model filter is designed, a multi-scale fusion feature extraction network is constructed, and the self attention module is connected to enhance the ability of the model to obtain the global information of the images and improve the detection performance; In order to reduce false detection rate, the authenticity judgement module is designed. The output heat map is used to judge whether the detected target is mistaken. Furthermore, a strategy of manipulating target source image to choose negative samples is applied to increase the distinguishing ability of model. The experiment result shows that the F1 score of improved RGB-N model is 0.759 on the data set with target stitching and erasure, the false detection rate is 0.2% on the non manipulated image data set, and the improved RGB-N model has a good robustness under JPEG compression attack.

Key words imaging systems; image manipulation detection; adaptive spatial rich model; multi-scale fusion; self-attention; authenticity judgement

OCIS codes 110.2970; 110.2960; 100.2000

收稿日期：2020-12-31；**修回日期**：2021-01-13；**录用日期**：2021-01-28

通信作者：*305240074@qq.com

1 引 言

随着图像处理技术的发展及图像处理软件的普及，图像的各种操纵及修改愈发简单，常人通过简单的操作说明即可使用功能强大的软件对图像进行修改，常用的修改方式有图像拼接（抠图）、目标擦除、目标仿制等。对于图像内容的操纵，一方面满足了各行业的业务需求，如平面设计、数码摄影后期的处理等；另一方面也可能被不法分子用于散播虚假消息、使用数字图像制作伪证，给社会带来危害。因此准确检测图像被操纵内容的位置及使用的手段，将为篡改图像行为的取证带来极大便利，防止虚假消息的大范围传播。

关于图像操纵检测与取证问题的研究已有多年的历史，根据所使用的取证手段，取证方法可大致分为主动取证和被动取证两类[1-3]。主动取证包括在制作数字图像时主动加入水印、标记等，适用于商业活动等需要主动验证图像真实性的场合；被动取证又称图像盲取证，旨在通过分析图像本身的特性而检测出图像是否经过操纵，无需提前植入验证信息，它的应用也更加广泛。因为图像操纵手段多种多样，且图像的来源不确定，所以对操纵图像进行被动取证具有很大的挑战性。文献[4-6]使用基于像素特性的方法进行取证。而随着 JPEG 图像格式的广泛使用，基于 JPEG 压缩特性的被动取证方法相继被提出。文献[7-8]利用 JPEG 图像双重压缩特性进行检测，文献[9]基于 JPEG 图像的块效应特征进行检测，文献[10-12]根据成像传感器本身的特性进行取证。

虽然上述方法能够取得一定的检测效果，但较为依赖图像单一属性，人工构造的特征只能适用于特定的场景，无法广泛应用。基于 JPEG 特征的检测方法对 JPEG 压缩更为敏感，图像如在传输过程中经过多次压缩则可能导致算法失效；而基于成像传感器特性的检测方法需要获知传感器的型号才能对其进行建模分析，适用性不强。近些年来，随着深度学习技术的迅速发展，Faster-RCNN[13]、YOLO[14]、SSD[15]等通用目标检测网络相继被提出，在检测速度与检测精度上均有不俗的表现。在图像操纵检测领域也开始引入深度学习技术，文献[16]使用卷积神经网络并行地提取图像的 RGB 特征及隐写特征，并利用融合特征检测图像中被操纵目标的位置及种类；文献[17]利用编码-解码器结构并结合 long short-term memory（LSTM）网络得到图像被操纵区域的像素级分割；文献[18]利用图像的元信息作为监督信号训练模型并判断图像内容的一致性。基于深度学习技术的图像操纵检测算法虽然适用性较强，不需要人工构造特征，但深度学习模型最终会学习到目标本身的语义信息，而忽略了目标在源图像和经操纵后的图像上的区别。因此深度学习模型在操纵数据集上有着较好的表现，而在未修改的正常图片上存在较高的误检率。

针对以上问题，本文基于 RGB-N 模型[16]设计了一种改进的图像操纵检测模型。首先描述了 RGB-N 模型的基本框架及自适应的空域富模型（SRM）滤波器；然后详细阐述了特征提取网络的设计细节，通过融合多尺度特征及自注意力模块增强网络对于图像上下文信息的提取能力，通过设计真实性判断模块和使用操纵图像-原始图像组成图像对的训练策略降低模型的误检率；最后通过实验对比，分析了所提算法与其他算法之间的效果差异，并通过消融实验验证了所提模型各模块产生的效果。

2 所提算法

为了能够有效地检测被操纵的图像内容，并降低误检率，提出了一种改进的 RGB-N 双流图像操纵检测网络，并制定了有针对性的训练及推理策略，基本结构如图 1 所示。

2.1 RGB-N 网络

RGB-N 是在 2018 年被提出的一种用于图像操纵检测的网络，基本结构如图 1 虚线部分所示。RGB-N 网络是基于 Faster-RCNN 架构的，除原有的 RGB 图像输入之外，又增加了噪声图像输入，噪声图像由 RGB 图像经过 SRM 滤波器转换后得到。SRM 滤波器的滤波核包括 KB 核、KV 核及二阶线性核，如图 2 所示。RGB 分支和噪声分支使用两个相同架构的骨干网络分别提取各自的图像特征，region proposal network（RPN）网络只使用 RGB 分支的特征进行训练，并选取感兴趣区域应用至 RGB 与噪声两个分支。选取的 RGB 特征用于目标框的回归，而 RGB 特征与噪声特征经过双线性池化操作后，得到的特征张量用于最终图像被操纵目标的分类。

原始 RGB-N 网络存在骨干网络特征提取能力有限及在正常图像上误检率较高的问题，针对以上问题对网络结构及训练策略进行了改进：1）使用了自适应的 SRM 滤波器，通过训练滤波器核中的少量参数，提高 SRM 滤波器在不同场景下的鲁棒

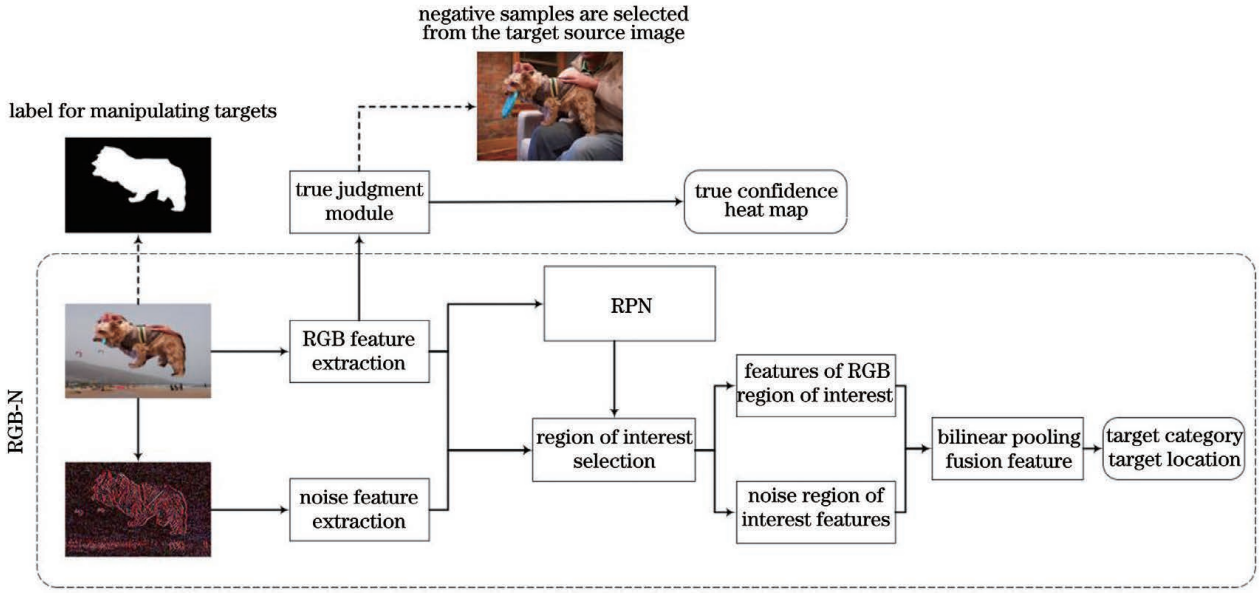

图 1　改进 RGB-N 模型结构

Fig. 1　Structure of improved RGB-N model

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

(a)　(b)　(c)

图 2　SRM 滤波器。(a)KB 核;(a)KV 核;(c)二阶线性核

Fig. 2　SRM filter. (a) KB kernel; (b) KV kernel; (c) second order linear kernel

性;2)设计多尺度融合的特征提取网络并引入自注意力模块来增强模型特征提取能力;3)设计真实性判断模块输出一个像素级别的热图,该热图代表每个点被操纵的概率,通过判断热图与模型目标框输出的一致性来降低误检率;4)训练模型时,负样本从操纵目标的源图像中选取,而不是从被操纵图像的背景中选取,使模型能学习到目标在源图像和被操纵后图像之间的区别,提高检测性能。

2.2　自适应 SRM 滤波器

SRM 是一种对图像残差进行建模的隐写分析方法,残差图像可抑制图像本身的内容,提高隐写特征的信噪比。对于图像 $I=(X_{i,j})\in\mathbf{R}^{n_1\times n_2}$, $i=1,\cdots,n_1$, $j=1,\cdots,n_2$,对应的残差图像为

$$Y=\mathrm{HP}(X)=X'_{i,j}-X_{i,j}\in\mathbf{R}^{n_1\times n_2},\quad(1)$$

式中:HP(X)表示残差图像 Y 为原图像 X 的高频部分;$X'_{i,j}$ 为图像中 $X_{i,j}$ 点处像素的邻域对 $X_{i,j}$ 的估计值。

RGB-N 选择了尺寸不同的 KB 核、KV 核及二阶线性核作为滤波器核提取图像的残差信息,为了保持输入网络前图像尺寸的一致性,将滤波核大小统一为 5×5,不足的部分进行补 0 处理。使用这种处理方式提取的残差信息对网络训练过程有一定的导向性,但滤波器所有值已固定,无法进一步优化,因此将图 2 中 KB 核及二阶线性核中的非 0 元素固定,扩展部分元素作为可训练的参数参与网络的训练过程,使模型能够根据不同场景自适应地调整滤波器部分数值。

2.3　特征提取网络

在真实检测场景中,图像中被操纵的区域尺寸往往具有较大的差距,如在进行图像拼接时,被操纵的目标可能在图像中尺寸占比很大,而在进行目标擦除操作时,目标的尺寸占比又很小,因此在提取特征时,需要充分保留各个尺度的信息。在 FPN[19] 和 Libra R-CNN[20] 基础上,设计了基于 bottom-up 和 top-down 结构的平衡特征多尺度特征提取网络,以获得更丰富的上下文信息,并根据噪声特征的特性改造了噪声提取网络中的池化层,以捕获较微弱的噪声信号。

2.3.1　多尺度特征融合

模型特征提取网络结构如图 3 所示,虚线部分

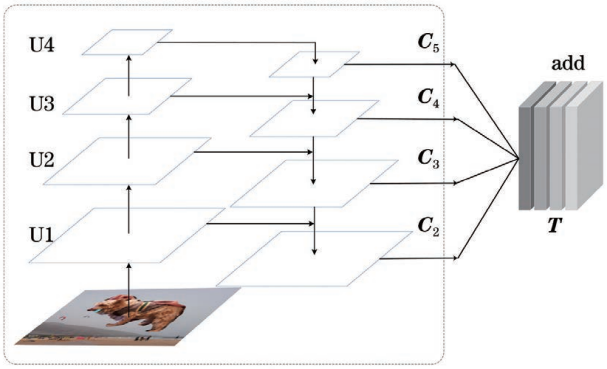

图 3　特征提取网络结构

Fig. 3　Structure of feature extraction network

为 bottom-up 和 top-down 结构，其中 $\boldsymbol{C}_i$ 代表图像的第 i 层特征，随着 i 的增加，特征图的语义信息逐渐增强，但空间信息逐渐减弱。为了更好地平衡各层级之间的特征信息，首先提取出不同层级的特征 $\{\boldsymbol{C}_2,\boldsymbol{C}_3,\boldsymbol{C}_4,\boldsymbol{C}_5\}$，并将所有层级的缩放映射到与层级 l 相同的尺寸，$l_{\max}=5$，$l_{\min}=2$，当 $\boldsymbol{C}_i>\boldsymbol{C}_l$ 时，对 $\boldsymbol{C}_i$ 进行最大池化，当 $\boldsymbol{C}_i<\boldsymbol{C}_l$ 时，则对 $\boldsymbol{C}_i$ 进行双线性插值。l 的取值不宜取最大值，否则具有高分辨率的特征图如 $\boldsymbol{C}_2$ 会由于过度池化而丢失细节信息；较小的 l 值会使高层特征图如 $\boldsymbol{C}_5$ 在插值过程中引入过多的噪声。因此目标层级宜选择中间层特征，l 设置为 4。当所有层级的特征都统一到同一尺度后，将它们融合就得到了多尺度融合特征张量 $\boldsymbol{T}$，表达式为

$$\boldsymbol{T}=\frac{1}{L}\sum_{l_{\min}}^{l_{\max}}\boldsymbol{C}_l。\tag{2}$$

2.3.2　自注意力模块

在卷积神经网络中，采用多次卷积-池化的操作逐步获得图像各部分的语义信息，但较小的卷积核得到的感受野有限，不能很好地利用图像的上下文信息。为了进一步提高网络对不同语义信息的区分度，使用自注意力模块[21]来获取图像各像素点之间的相关性，提高网络特征提取能力。

自注意力模块结构如图 4 所示，在得到融合特征 $\boldsymbol{T}$ 后，分别使用 3 个 1×1 卷积核对 $\boldsymbol{T}(B,H,W,C)$ 进行卷积，将通道数由 C 降为 $C/8$，其中 B 为批数量，H 和 W 为特征图的高和宽。f，g，h 三个分支经过 1×1 卷积后的输出分别为 $\boldsymbol{T}_{\mathrm{f\_o}}(B,H,W,C/8)$，$\boldsymbol{T}_{\mathrm{g\_o}}(B,H,W,C/8)$，$\boldsymbol{T}_{\mathrm{h\_o}}(B,H,W,C/8)$；将 $\boldsymbol{T}_{\mathrm{f\_o}}$、$\boldsymbol{T}_{\mathrm{h\_o}}$ 除通道外的三维合并后得到 $\boldsymbol{T}_{\mathrm{f\_o\_c}}(BHW,C/8)$、$\boldsymbol{T}_{\mathrm{h\_o\_c}}(BHW,C/8)$；将 $\boldsymbol{T}_{\mathrm{g\_o}}$ 通道合并后进行转置，得到 $\boldsymbol{T}_{\mathrm{g\_o}}^{\mathrm{T}}(C/8,BHW)$；将 $\boldsymbol{T}_{\mathrm{f\_o\_c}}$ 和 $\boldsymbol{T}_{\mathrm{g\_o}}^{\mathrm{T}}$ 作内积并经过 sotfmax 层即得到了各像素点之间的相关性矩阵，该矩阵代表了整个特征图上每个点对于其他特征点的影响，引入了图像的上下文信息，再将该矩阵与 $\boldsymbol{T}_{\mathrm{h\_o}}$ 作内积即可得到自注意力特征；最后使用 1×1 卷积将自注意力特征的通道数由 $C/8$ 恢复至 C，再与原始融合特征 $\boldsymbol{T}$ 作残差运算，即得到最终的输出特征 $\boldsymbol{P}$。

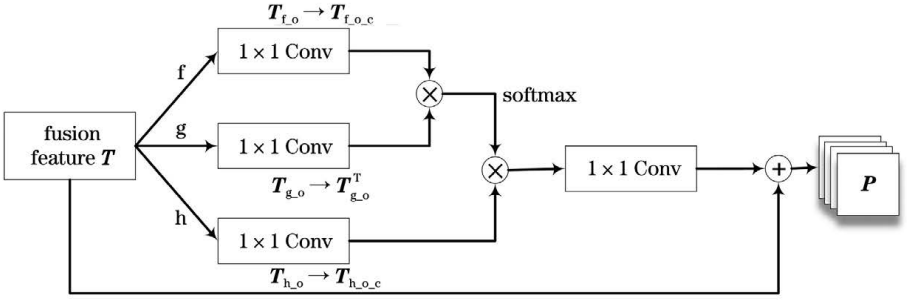

图 4　自注意力模块

Fig. 4　Self-attention module

在得到自注意力特征图后，按照与 2.3.1 节多尺度特征融合中所述相反的操作将特征分别还原至 $\{\boldsymbol{C}_2,\boldsymbol{C}_3,\boldsymbol{C}_4,\boldsymbol{C}_5\}$ 的原始尺度，得到最终的特征输出 $\boldsymbol{P}=\mathrm{Concat}(\boldsymbol{P}_2,\boldsymbol{P}_3,\boldsymbol{P}_4,\boldsymbol{P}_5)$。

2.3.3　噪声特征提取

卷积-激活-池化的结构通常能够较好地提取图像的语义信息，其中池化的方式包括最大池化和平均池化两种，表达式分别为

$$\mathrm{max\_pooling}(A_m)=\max_{\alpha_n\in A_m}\alpha_n,\tag{3}$$

$$\mathrm{average\_pooling}(A_m)=\frac{1}{|A_m|}\sum_{\alpha_n\in A_m}\alpha_n,\tag{4}$$

式中：A_m 为特征图的第 m 个池化区域；α_n 为该区域的第 n 个元素。噪声图中的残差信号通常不够明显，如果在网络的底层使用最大池化方式处理特征图极容易造成关键信息的丢失，平均池化则能融合更多区域内的信息，对噪声特征的提取更加有利，因

此改变噪声特征提取网络不同特征层的池化方式可以提升模型的检测性能。

2.4　真实性判断模块

RGB-N 模型虽然在使用拼接、复制-粘贴等手段制作的图像操纵数据集上取得不错的检测效果，但在未经任何修改的原始图像上具有较高的误检率，而实际的检测场景往往是大量真实图片中混有少量被操纵的图像，这极大地限制了 RGB-N 模型的具体应用。为此设计了一种用于分辨真实图片与操纵图片的真实性判断模块，以降低所提模型对于真实图像的误检率。

真实性判断模块结构如图 5 所示，输入为 RGB 分支融合后的特征 $\boldsymbol{P}$，使用 SE 模块[22]来增强网络的注意力，令缩放参数 $r=8$，通道数 C 与特征 $\boldsymbol{P}$ 的通道数保持一致，设置为 512；之后接入 $3\times3\times(C/4)$、$3\times3\times N$ 的两个卷积层，其中 N 为操纵图像方式的数量，如本实验中可检测的图像操纵方式包括图像拼接、图像擦除两种，因此 $N=2$。特征图经过 softmax 激活后，最终得到一个像素级别的热图，热图中每个像素的值代表不同种类的分类概率。训练模型时的正样本代表存在操纵痕迹的图像中被操纵的部分，负样本代表存在操纵痕迹的图像中未被操纵的部分。训练模型时，把操纵图像的背景和未操纵原始图像的全部作为负样本，这样正样本的占比会远小于负样本，并通过组合 Dice 损失函数[23]和交叉熵损失函数（BCE）的方式处理正负样本分布不均衡的情况。真实性判断模块的损失函数为

$$L_{\text{heat}}=-\frac{1}{N'}\sum_{o}^{N'}\left[y_o\log p_o+(1-y_o)\log(1-p_o)\right]+1-\frac{2\sum_{o}^{N'}p_o y_o}{\sum_{o}^{N'}p_o^2+\sum_{o}^{N'}y_o^2},\tag{5}$$

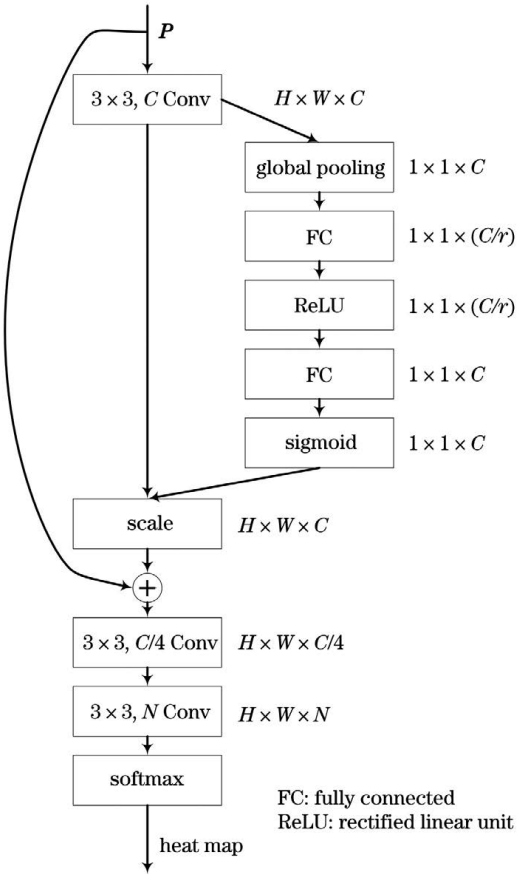

图 5　真实性判断模块

Fig. 5　Authenticity judgement module

式中：N' 为操纵图像与目标来源原始图像特征图所有点集合的个数；p_o 为模型输出的第 o 个特征点；y_o 为第 o 个特征点的真实标注。

所提模型的最终损失函数为

$$L=L_{\text{cls}}+L_{\text{reg}}+\lambda_1(L_{\text{RPN\_cls}}+L_{\text{RPN\_reg}})+\lambda_2 L_{\text{heat}}。\tag{6}$$

式中：$L_{\text{RPN\_cls}}$、L_{cls} 为 RPN 部分的分类损失函数、模型最终的分类损失函数；$L_{\text{RPN\_reg}}$、L_{reg} 为 RPN 部分的框回归损失函数、模型最终的框回归损失；L_{heat} 为真实性判断模块的损失函数；λ_1 与 λ_2 分别为 RPN 部分损失的权重和真实性判断损失的权重。模型的分类损失均使用交叉熵损失函数，框回归损失均使用 smooth L_1 损失函数。

2.5　实现细节

2.5.1　网络参数

所提网络可实现端到端的训练。训练集中所有的图像均为 JPEG 格式图像，且对其中一半以 70 的质量因数进行 JPEG 压缩，另一半不压缩，图像输入网络之前调整为短边不小于 600 pixel，长边不大于 1000 pixel 并进行水平翻转以扩充训练集。基础锚框的尺寸为 8^2，16^2，32^2，64^2，长宽比为 1∶2，1∶1，2∶1。优化器为 Adam，一阶矩估计的指数衰减率 $\beta_1=0.9$，二阶矩估计的指数衰减率 $\beta_2=0.999$，$\varepsilon=10^{-8}$，采用余弦衰减学习率及 warm_up 学习率调整

策略，学习率在训练的初始阶段由 0 增加至 0.001，当 epoch 大于 10 时，开始衰减。RPN 部分正样本的阈值设置为大于 0.7，负样本的阈值设置为小于 0.3，非极大值抑制(NMS)的阈值设置为 0.2。

2.5.2 训练样本的标注与选择

训练模型需要输入的数据共有 3 部分：1)含真实框坐标标注的被操纵图像；2)被操纵区域边界向外扩展后的掩模；3)被操纵目标来源的原始图像。

常规目标检测网络需要学习的是目标与背景之间的区别，而在图像被操纵目标的检测中不仅需要学习目标与背景之间的区别，更需要网络能够识别出未操纵目标与被操纵目标之间的区别，因此所提模型使用操纵目标的原始图像与被操纵后的图像组成图像对的形式进行训练，模型训练的正样本从被操纵图像中选择，而负样本则从原始图像中选择，如图 6 所示。如果使用图 6(a)的操纵目标作为正样本，背景作为负样本，则模型无法明显地区分不同背景下的同一目标是否经过拼接处理；如果正样本选取方式不变，而负样本从图 6(b)中选取，使得模型学习到被拼接的目标与真实场景中的目标的区别。而在真实性判断网络中正样本为被操作目标区域，负样本为被操纵图像非目标部分与整个原始图像。图 6(c)为图像操纵区域的像素级标注。

图 6 正负样本选取方式示意图。(a)拼接后的图像；(b)拼接目标来源图像；(c)操纵目标标签

Fig. 6 Schematic diagram of positive and negative sample selection method. (a) Spliced image; (b) source image of mosaic target; (c) manipulation target label

2.5.3 推理过程

在模型执行推理时，需要综合考虑主网络输出的目标回归框与注意力网络输出的热图，其中被操纵区域及类型主要由主网络输出得到，而真实性判断模块则用于辅助判断主网络输出是否为误检。具体策略为：将主网络输出的目标框缩放至与热图同一分辨率，目标框区域为矩形，而热图中的正样本区域形状往往是不规则的，因此为了使真实性判断的热图能对目标框输出进行二次校正，先选取热图中包含在目标框内置信度大于 0.5 的点，以此来确定输出框与热图的交集中被判断为操纵区域的像素点，再计算这些点置信度均值，若均值大于最终的分类置信度阈值 0.8，才认为该目标是被操纵的。

3 实验与结果分析

3.1 数据集和评估指标

深度学习模型的训练需要大量数据，而现实中难以收集到如此数量的操纵图像数据，因此所提模型采用自动合成的方式制作操纵图像数据集，而操纵对象的基础图像则来源于 common objects in context(COCO)数据集。COCO 数据集是微软公司于 2014 年提供的一个用于图像识别的数据集，并为图像提供了包括实例分割、目标关键点等详细标注。

以 COCO 数据集为基础，设计了操纵图像的自动化生成方法，并制作了含 6 万张操纵图像的数据集，其中以图像拼接方式制作的图像有 3 万张，以目标擦除方式制作的图像有 3 万张，操纵图像数据集以 9∶1的比例分为训练集和测试集。

3.1.1 图像拼接

图像拼接利用 COCO 数据集提供的实例分割注释随机截取一张图中的目标粘贴至另一张图像中，并以 30%概率对图像进行高斯模糊处理，降低裁剪目标边缘与背景之间的明显程度。最终生成的数据为拼接图像、原始背景图像组成的图像对，同时记录裁剪目标在背景图上的坐标位置及实例标注，如图 7(a)所示。

3.1.2 目标擦除

为了提高目标擦除后的图像质量，将图像中的目标擦除后，再采用文献[24]所述方法对图像进行修复，以保证擦除区域与背景的一致性。最终生成擦除-修复后的图像与原始图像组成的图像对及相应的标注，如图 7(b)所示。

使用 F1 分数作为考量模型对于被操纵目标检测能力的评价指标，如文献[25]所述，对于每一张被

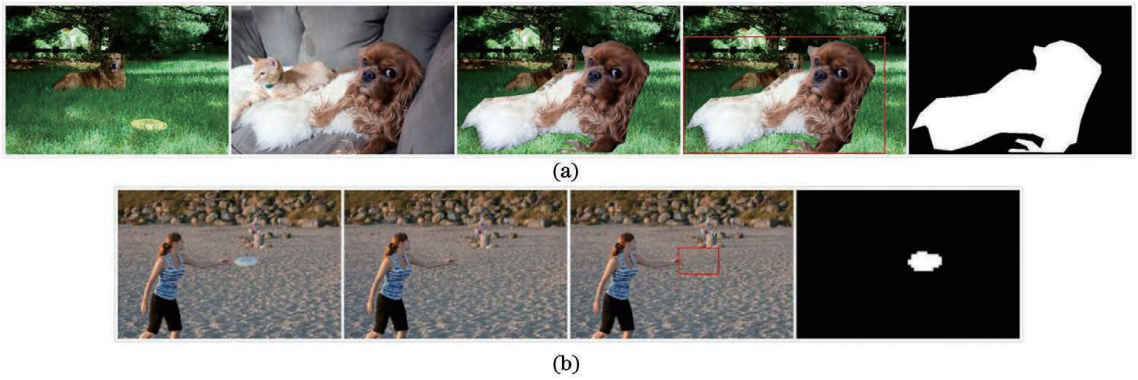

图 7　训练样本及标注样本。(a)目标拼接;(b)目标擦除与修复

Fig. 7　Training samples and labeled samples. (a) Target splicing; (b) target erasure and repair

操纵的图像,把真实操纵区域作为正样本,而未操纵部分作为负样本,分别计算单张图像上像素级别的召回率 R 和精确率 P,并计算该图像的 F1 分数,最后计算测试集所有图像的均值作为模型最终的 F1 分数。单张图像的 F1 分数计算公式为

$$S_{\mathrm{F1}}=\frac{2\times P\times R}{P+R}。\tag{7}$$

F1 分数虽然同时考虑了图像的召回率和精确率,但该指标无法反映检测模型在没有任何操纵区域的正常图像上出现误报警的情况,因此引入了误检率 M 这一指标,并从 COCO 数据集中选取 2000 张未修改的图像作为测试集,若在正常图像上检测到了操纵区域,则认为该图像出现了误检,最终统计误检图像在正常图像测试集中的占比作为误检率指标。

$$M=\frac{N_{\mathrm{M}}}{N_{\mathrm{T}}}\times 100\%,\tag{8}$$

式中:N_{T} 为测试集图像总数量;N_{M} 为测试集中提示存在操纵痕迹图像的数量。

3.2　实验结果

对比了所提算法与 ELA[26]、DCT[7]、NOI1[27]、Faster-RCNN[13]、EXIF-Consistency[18]、RGB-N[16]等基准模型的实验效果。其中 ELA 通过被操纵区域与真实区域之间不同的 JPEG 压缩误差来找到操纵目标区域,DCT 通过分析图像在 JPEG 重压缩过程中的块效应定位被操纵区域,NOI1 对图像局部噪声进行建模并通过噪声不一致性来找出操纵位置。Faster-RCNN、EXIF-Consistency、RGB-N 均为基于深度学习的方法,Faster-RCNN 为一种经典的二阶段目标检测框架,EXIF-Consistency 是一种基于图像 EXIF 元数据的图像拼接检测方法,RGB-N 为一种融合图像 RGB 特征及隐写特征的双流检测网络。为确保实验的公平性,基于开源代码复现了上述算法,且保证各算法在原论文中所使用的测试集上达到相同或更优的结果。实验中 F1 分数均在操纵图像数据集进行测试,误检率均在未操纵的正常图像进行测试。

最终的实验结果如表 1 所示,可见虽然 ELA、DCT、NOI1 算法整体误检率较低,但在操纵图像测试集上表现不佳,因为它们仅关注图像被操纵后留下的特定特征,从而限制了算法的性能。Faster-RCNN 的 F1 分数相较于传统算法有了较大的提升,但在训练过程中模型主要学习到了目标的语义信息,没有区分普通目标与操纵后目标的能力;EXIF-Consistency 与 RGB-N 分别引入图像元信息、图像隐写特征作为监督条件辅助深度学习模型进行训练,RGB-N 在操纵图像数据集上有更好的表现,而 EXIF-Consistency 的误检率更低。所提算法在 F1 分数和误检率两个评价指标上均表现较好。

表 1　各算法的测试结果

Table 1　Results of each algorithm

Algorithm	F1 score	M /%
ELA	0.235	1.8
DCT	0.434	2.3
NOI1	0.287	1.9
Faster-RCNN	0.570	16.1
EXIF-Consistency	0.683	3.7
RGB-N	0.722	16.7
Proposed algorithm	0.759	0.2

由于 JPEG 格式在文件压缩上的灵活性,在手机拍照等真实场景中绝大部分图像采用该格式进行保存和传输,因此算法在 JPEG 压缩操作下的鲁棒性尤为重要,为此对比测试了各算法在图像经过 JPEG 不同品质因数(QF)压缩后的检测结果。实

验结果如表 2 所示，DCT 在 JPEG 压缩攻击下趋于失效，ELA、NOI1 算法同样受到了较大影响，而所提算法在不同品质因数压缩攻击下的表现均优于其他 6 种算法。

表 2　各算法在使用不同品质因子压缩时的 F1 分数

Table 2　F1 score of each algorithm when using different quality factors for compression

Algorithm	*QF* 100	*QF* 90	*QF* 80	*QF* 70	*QF* 60	*QF* 50
ELA	0.235	0.231	0.229	0.223	0.215	0.207
DCT	0.434	0.205	0.198	0.185	0.103	0.096
NOI1	0.287	0.285	0.281	0.274	0.258	0.235
Faster-RCNN	0.570	0.570	0.567	0.564	0.559	0.550
EXIF-Consistency	0.683	0.678	0.677	0.671	0.661	0.653
RGB-N	0.722	0.722	0.719	0.716	0.713	0.708
Proposed algorithm	0.759	0.759	0.759	0.754	0.742	0.738

3.3　消融实验

3.3.1　自适应 SRM 滤波器

针对 SRM 滤波器扩展部分参数训练的实验，分别采取如下方式训练模型：1）RGB-N 网络直接补 0；2）二阶线性核补 0，只训练 KB 核扩展部分参数；3）KB 核补 0，只训练二阶线性核扩展部分参数；4）同时训练 KB 核、二阶线性核扩展部分参数。为公平地对比各方法的效果，在使用不同方法训练 RGB-N 网络时均迭代 30 万次，每隔 2 万次保存模型，从中选择 F1 分数最高的作为该方法的基准模型。实验中各算法均在迭代 8 万次时具有最高的 F1 分数，随着训练轮数的增加，各模型均出现了过拟合现象，检测性能开始下降。迭代 8 万次时的实验结果如表 3 所示，从表中可以看出，只训练 KB 核时获得了最好的结果，训练二阶核或同时训练 KB 与二阶核时会得到更差的效果，说明在噪声图像提取阶段如果训练太多的参数会使预处理层失去原有的导向性；而固定大部分参数，只训练 KB 核扩展的少量参数，能够使滤波器在保持原有噪声特征提取能力的基础上根据不同的场景进行针对性的优化，提高了鲁棒性。各方法提取的噪声图像如图 8 所示，可见只训练 KB 核扩展部分参数时具有最好的噪声提取效果。

表 3　不同滤波器参数参与训练的实验结果

Table 3　Results of training different filter parameters

Training kernel	F1 score	*M* /%
None	0.718	17.9
KB kernel	0.722	16.7
Second order kernel	0.684	18.5
KB and second order kernel	0.659	19.3

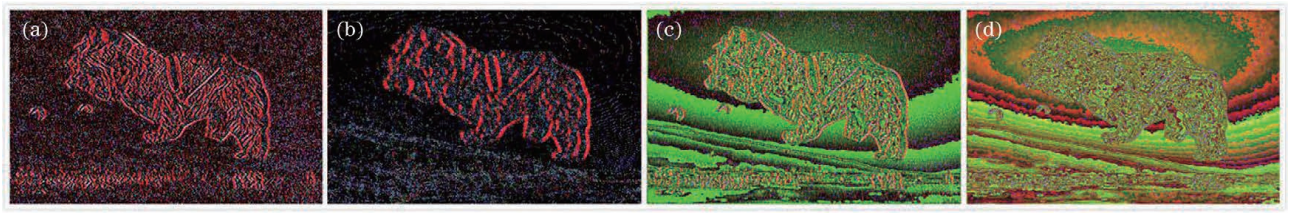

图 8　训练不同滤波器参数的可视化效果。(a)不训练滤波器；(b)KB 核；(c)二阶线性核；(d)KB、二阶线性核

Fig. 8　Visual effects of training different filter parameters. (a) None; (b) KB kernel; (c) second order linear kernel; (d) KB kernel and second order linear kernel

3.3.2　网络部分

为了分析网络每个模块及训练策略对最终结果的贡献，以加入自适应 SRM 滤波器核的 RGB-N 网络为基础，逐步增加自注意力特征提取网络、真实性判断模块及正负样本选择策略，整体测试结果如表 4 所示。消融实验除需验证的模块、策略外其余参数保持一致，模型的骨干网络均采用 ResnNet-101 FPN，以保证各部分实验结果的准确性。

表 4　消融实验结果

Table 4　Results of ablation experiment

	Self-attention module	Average pooling	Authenticity judgement module		Sample selection	F1 score	M /%
			Loss	Loss and output			
Faster-RCNN						0.570	16.1
					√	0.572	11.6
Proposed algorithm						0.722	16.7
	√					0.753	16.9
	√	√				0.754	16.0
	√	√	√			0.759	11.5
	√	√	√	√		0.758	3.5
	√	√	√	√	√	0.759	0.2

3.3.2.1　自注意力特征提取网络

自注意力特征提取网络加入后，与原始 ResnNet-101 FPN 骨干网络相比，所提算法的 F1 分数提高了 0.031，说明不同尺度特征图之间的融合及自注意力模块明显提高了模型对操纵目标的辨识能力，但同时模型整体的误检率也由 16.7%增加至 16.9%。

3.3.2.2　噪声特征提取

表 5 对比了在使用自注意力特征提取网络时改变噪声网络不同特征层的池化方式的实验结果。从表中可以看出，在 U1 与 U2 层使用平均池化具有最好的效果，在更高层网络加入平均池化会使网络整体性能变差，说明在底层网络使用平均池化可更好地捕获残差信息，但在高层网络使用平均池化反而会降低网络的性能。

表 5　噪声网络池化方式实验结果

Table 5　Results of noise network pooling method

Average pooling layer	F1 score	M /%
None	0.753	16.9
U1	0.754	16.1
U1＋U2	0.754	16.0
U1＋U2＋U3	0.746	17.1
U1＋U2＋U3＋U4	0.741	17.8

3.3.2.3　真实性判断模块

原始 RGB-N 模型在设计时并未对检测原始图像时的误检率进行更多考虑，在原始 COCO 数据集检测上的误检率高达 16.7%，且随着特征提取能力的增强，误检率也随之提高。在加入真实性判断模块后分别使用两种策略调整输出结果：1)在训练时，加入热图损失，推理时不使用输出热图，误检率由 16.9%下降至 11.5%，F1 分数由 0.753 增加至 0.759；2)在训练时加入，推理时使用 2.5.3 节所述方法结合热图得到最终输出，误检率下降至 3.5%，F1 分数为 0.758。可见真实性判断分支的加入可以在一定程度帮助模型在训练阶段提高对操纵图片和真实图片的辨别能力，结合最终输出的热图可以很好地控制误检率。

以加入自注意力模块、噪声特征平均池化、真实性判断模块及样本选择策略后的改进网络为基准，在真实性判断网络分支中使用不同损失函数的实验结果如表 6 所示。从表中可以看出，交叉熵损失函数会因为正样本数量过少导致模型检测性能下降，Dice loss 会明显改善样本不均衡的问题，交叉熵与 Dice loss 结合的损失函数可获得最佳检测效果。

表 6　不同损失函数的实验结果

Table 6　Results of different loss functions

Loss function	F1 score	M /%
BCE	0.753	0.8
Dice loss	0.757	0.2
BCE＋Dice Loss	0.759	0.2

3.3.2.4　训练样本选择策略

将负样本由从被操纵图像背景中选择调整为从操纵目标来源图像全图中选择后，模型的误检率最终降低至 0.2%，同时对 F1 分数没有明显的影响。不失一般性，在 Faster-RCNN 模型中采用同样的样本选择策略，F1 分数由 0.570 提升至 0.572，误检率由 16.1%下降至 11.6%，说明该策略的有效性，且可以方便地应用至其他基于通用目标检测框架的图像操纵检测模型中。

3.4 定性结果

图 9 显示了模型在各阶段的可视化效果，其中图 9(a)、图 9(b)为分别使用目标拼接及目标擦除与修复方式操纵图像。可见被操纵的图像区域在噪声图与真实性模块输出的热图里均留下了明显的痕迹，最终模型成功检测到了被操纵的目标位置及种类。在图 9(c)中，输入模型的是未经操纵的正常图像，其中的人物目标被模型的 RGB-N 部分误识别为被拼接的目标，而在真实性热图中并没有提示存在操纵区域，根据判定准则，模型最终没有进行提示，避免了误检的情况。

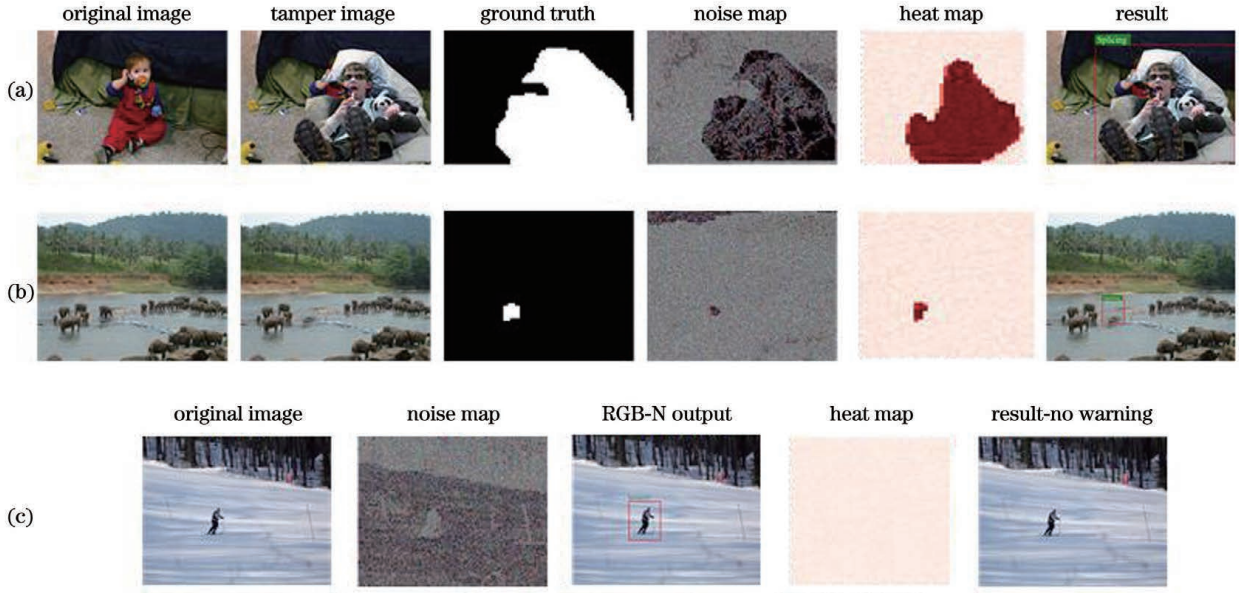

图 9　模型可视化输出。(a)目标拼接；(b)目标擦除与修复；(c)正常图像

Fig. 9　Visual outputs of model. (a) Target splicing; (b) target erasure and repair; (c) normal images

4 结　论

提出了一种基于改进 RGB-N 的图像操纵检测模型。所提模型通过融合多尺度特征来获取更丰富的网络浅层信息，使用自注意力模块增强图像特征的内在相关性，提高了对图像中被操纵目标的检测能力；设计了真实性判断模块并采用从操纵目标来源图像获取负样本的训练策略降低了模型的误检率。实验结果验证了所提算法的有效性，并通过消融实验验证了不同模块的具体作用。在未来的工作中，将对图像操纵方式进行进一步细分，并探究针对不同操纵方式更加鲁棒的图像特征。

参　考　文　献

[1] Wang J, Zhang Y C, Huo Z Q, et al. Image tampering detection method based on approximate nearest neighbor search[J]. Laser & Optoelectronics Progress, 2020, 57(10): 101102.
王静, 张雨辰, 霍占强, 等. 基于近似最近邻搜索的图像篡改检测方法[J]. 激光与光电子学进展, 2020, 57(10): 101102.

[2] Liu T T, Zhang Y J, Wu F, et al. Diffusion-based image inpainting forensics via gradient domain guided filtering enhancement[J]. Laser & Optoelectronics Progress, 2020, 57(8): 081003.
刘婷婷, 张玉金, 吴飞, 等. 基于梯度域导向滤波增强的图像扩散修复取证[J]. 激光与光电子学进展, 2020, 57(8): 081003.

[3] Guo J C, Wei H W, He Y H, et al. Enhancing image steganographic security using daubechies wavelet[J]. Laser & Optoelectronics Progress, 2019, 56(3): 031004.
郭继昌, 魏慧文, 何艳红, 等. 使用 Daubechies 小波增强图像隐写安全性[J]. 激光与光电子学进展, 2019, 56(3): 031004.

[4] Fridrich J, Soukal D, Jan Lukáš. Detection of copy-move forgery in digital images[EB/OL]. [2020-12-25]. https://ia.binghamton.edu/publication/FridrichPDF/copymove.pdf.

[5] Ng T T, Chang S F, Sun Q B. Blind detection of photomontage using higher order statistics[C]//2004 IEEE International Symposium on Circuits and Systems, May 23-26, 2004, Vancouver, BC, Canada. New York: IEEE Press, 2004: V-688-V-691.

[6] Fu D D, Shi Y Q, Su W. Detection of image splicing based on Hilbert-Huang transform and moments of

characteristic functions with wavelet decomposition [M] // Shi Y Q, Jeon B. Digital watermarking. Lecture notes in computer science. Heidelberg: Springer, 2006, 4283: 177-187.

[7] Bianchi T, Piva A. Image forgery localization via block-grained analysis of JPEG artifacts [J]. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 1003-1017.

[8] Bianchi T, Piva A. Detection of nonaligned double JPEG compression based on integer periodicity maps [J]. IEEE Transactions on Information Forensics and Security, 2012, 7(2): 842-848.

[9] Fan Z G, de Queiroz R L. Identification of bitmap compression history: JPEG detection and quantizer estimation[J]. IEEE Transactions on Image Processing, 2003, 12(2): 230-235.

[10] Johnson M K, Farid H. Exposing digital forgeries through chromatic aberration [C] // Proceeding of the 8th workshop on Multimedia and security-MM & Sec'06, September 26-27, 2006, Geneva, Switzerland. New York: ACM Press, 2006: 48-55.

[11] Swaminathan A, Wu M, Liu K J R. Nonintrusive component forensics of visual sensors using output images[J]. IEEE Transactions on Information Forensics and Security, 2007, 2(1): 91-106.

[12] Chen M, Fridrich J, Goljan M, et al. Determining image origin and integrity using sensor noise [J]. IEEE Transactions on Information Forensics and Security, 2008, 3(1): 74-90.

[13] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[14] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 779-788.

[15] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[M] //Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[16] Zhou P, Han X T, Morariu V I, et al. Learning rich features for image manipulation detection [C] //2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 1053-1061.

[17] Bappy J H, Simons C, Nataraj L, et al. Hybrid LSTM and encoder-decoder architecture for detection of image forgeries [J]. IEEE Transactions on Image Processing, 2019, 28(7): 3286-3300.

[18] Huh M, Liu A, Owens A, et al. Fighting fake news: image splice detection via learned self-consistency[M] //Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 9905: 21-37.

[19] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 936-944.

[20] Pang J M, Chen K, Shi J P, et al. Libra R-CNN: Towards balanced learning for object detection [C] // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 821-830.

[21] Wang X L, Girshick R, Gupta A, et al. Non-local neural networks[C] //2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7794-7803.

[22] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.

[23] Milletari F, Navab N, Ahmadi S A. V-net: fully convolutional neural networks for volumetric medical image segmentation [C] // 2016 Fourth International Conference on 3D Vision (3DV), October 25-28, 2016, Stanford, CA, USA. New York: IEEE Press, 2016: 565-571.

[24] Nazeri K, Ng E, Joseph T, et al. EdgeConnect: generative image inpainting with adversarial edge learning [EB/OL]. (2019-01-01) [2020-12-25]. https://arxiv.org/abs/1901.00212.

[25] Salloum R, Ren Y Z, Jay Kuo C C. Image splicing localization using a multi-task fully convolutional network (MFCN)[J]. Journal of Visual Communication and Image Representation, 2018, 51: 201-209.

[26] Krawetz N, Solutions H F. A picture's worth [J]. Hacker Factor Solutions, 2007, 6(2): 2-31.

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