面向强后处理场景的图像篡改定位模型

谭舜泉[1,2,3]，廖桂樱[1,2,3]，彭荣煊[2,3,4]，黄继武[5]

（1.深圳大学计算机与软件学院，广东 深圳518060；2.深圳市媒体信息内容安全重点实验室，广东 深圳518060；
3.广东省智能信息处理实验室，广东 深圳518060；4.深圳大学电子与信息工程学院，广东 深圳 518060；
5.深圳北理莫斯科大学工程系智能感知与计算广东省重点实验室，广东 深圳 518116）

摘　要：针对微信、微博等社交平台对图像进行的压缩、尺度拉伸等有损操作带来的篡改痕迹模糊或被破坏的挑战，提出了一种对抗强后处理的图像篡改定位模型。该模型选用了基于Transformer的金字塔视觉转换器作为编码器，用于提取图像的篡改特征。同时，设计了一个类UNet结构的端到端编码器-解码器架构。金字塔视觉转换器的金字塔结构和注意力机制可以灵活关注图像的各个区块，结合类UNet结构能够多尺度地提取图像上下文间的关联信息，对强后处理的图像有着较好的鲁棒性。实验结果表明，所提模型在对抗JPEG压缩、高斯模糊等常见的后处理操作以及在不同社交媒体传播场景的数据集上的定位性能上明显优于目前主流的篡改定位模型，展现出了优异的鲁棒性。

关键词：强后处理场景；图像篡改定位；鲁棒性；金字塔视觉转换器

中图分类号：TN391

文献标志码：A

DOI：10.11959/j.issn.1000−436x.2024079

Image tampering localization model for intensive post-processing scenarios

TAN Shunquan[1,2,3], LIAO Guiying[1,2,3], PENG Rongxuan[2,3,4], HUANG Jiwu[5]

1. College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
2. Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen Key Laboratory of Media Security, Shenzhen 518060, China
3. Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen 518060, China
4. College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China
5. Guangdong Laboratory of Machine Perception and Intelligent Computing, Faculty of Engineering, Shenzhen MSU-BIT University, Shenzhen 518116, China

Abstract: Addressing the challenges of blurred or destroyed tampering traces presented by lossy operations such as image compression and scaling on images within social platforms like WeChat and Weibo, an adversarial image tampering localization model was introduced. Utilizing the pyramid vision transformer, which was built upon the Transformer architecture, as an encoder for extracting tampering features from images. Simultaneously, an end-to-end encoder-decoder structure, reminiscent of the UNet architecture, was formulated. The pyramid structure and attention mechanisms inherented to the pyramid vision transformer afforded a flexible examination of diverse image regions. When integrated with the UNet-like architecture, it facilitated multiscale contextual information extraction, thereby fortifying the model's resilience to intense post-processing effects. Empirical results illustrate that the proposed model exhibits a substantial performance advantage over conventional tampering localization models, particularly in scenarios involving prevalent post-processing techniques such as JPEG compression and Gaussian blur. Notably, the model demonstrates exceptional robustness in assessments conducted with datasets representing diverse social media dissemination scenarios.

Keywords: intensive post-processing scenario, image tampering localization, robustness, pyramid vision transformer

---

收稿日期：2023−11−13；**修回日期：**2024−03−11

通信作者：黄继武，jwhuang@szu.edu.cn

基金项目：国家自然科学基金资助项目（No.62272314, No.U23B2022）；广东省重点实验室基金资助项目（No.2023-B1212060076）

Foundation Items: The National Natural Science Foundation of China (No.62272314, No.U23B2022), Guangdong Provincial Key Laboratory (No.20231B1212060076)

0　引言

随着数字图像处理技术的不断发展与普及，即使是缺乏专业知识的个体，也能制作逼真的篡改图像。若这些伪造图像在网络上被滥用传播，则会对公共安全等造成严重威胁，甚至危及国家安全[1]。因此，针对图像篡改的定位算法具有重要的应用价值。

目前存在的篡改方式主要有：拼接，将 2 个或多个不同的图像片段拼接在一起；复制移动，将图像内的某一部分复制并移动到其他位置；删除，删除或消除图像中特定的区域。

篡改技术的不断发展使篡改图像的视觉系统变得难以分辨，因此设计相关的方法来定位篡改区域变得至关重要。从机理上看，对原图像进行篡改操作会留下相关的痕迹，而这些痕迹是完成图像篡改检测或定位任务的重要判断依据。在早期的图像篡改定位工作中，主要采用手工设计的特征来实现篡改定位任务。例如，Bianchi 等[2]提出了一种方法，用于在 8×8 的图像块中分析 JPEG 格式重压缩产生的伪影，以定位篡改区域。Chierchia 等[3]采用贝叶斯框架来估计传感器模式噪声，并利用马尔可夫随机场（MRF, Markov random field）对像素间的关系进行建模，从而预测篡改区域。此外，Korus 等[4]通过将多种尺寸滑动窗口的输出结果融合在一起，提高了篡改定位的性能。然而这些利用手工设计特征的定位算法受限于设计者的经验和知识；同时，它们大多针对某一种篡改手段进行定义，并且在大尺寸图像上的滑动窗口定位方式耗费过多的计算资源[5]。因此，面对复杂的篡改操作和多样的图像来源，传统的篡改定位方法无法很好地发挥作用。

近年来，深度学习在多个领域取得显著成就。应用适当的深度学习模型架构，可以构建端到端的图像篡改定位模型，从图像中自动学习并提取有效的篡改特征。目前，基于深度学习的图像篡改定位模型已经成为主流的方法。例如，RGB-N[6]基于 R-CNN（region-convolutional neural network）架构，提出一种双流 Faster R-CNN，其中，第一个流用于提取 RGB 特征并识别篡改痕迹，而第二个流则利用噪声特征来建模篡改区域与真实区域之间的噪声差异，从而增强定位篡改区域的准确性。Dense-FCN[7]则采用了全卷积网络（FCN, fully convolutional network）结构，设计了一种由密集连接和空洞卷积构成的全卷积编码器-解码器架构，以提升篡改定位性能。ManTra-Net[8]利用了 CNN 和长短期记忆（LSTM, long short-term memory）网络架构，设计了一个端到端的网络，能够同时进行检测和定位。该方法将问题视为异常检测，并引入了 LSTM 来评估局部异常情况。此外，MVSSNet[9]通过多视角特征学习和多尺度监督 2 个支路，学习对篡改操作敏感的语义无关特征，从而提高伪造区域定位的准确性。SAT-IFL[10]则采用对抗性训练，利用由 FGSM[11]生成的对抗样本进行训练，以提升模型定位的性能。IF-OSN[12]模型通过分别对线上社交平台上引入的可预测噪声和不可见噪声进行建模，能够有效地检测社交网络上的篡改图像。Guo 等[13]为篡改或合成图像定义了分层细粒度标签，使检测器不仅可以学习综合特征，还可以学习不同属性的固有层次性质。然而，当训练集多样性受限时，这种方法可能面临挑战。文献[14]提出了一种 JPEG 压缩特征提取器，并利用自监督学习策略，即便在有限的训练数据下也能有效定位不同 JPEG 压缩处理的图像区域。随着生成模型（如扩散模型 Diffusion）的发展，当前不少研究[15-17]是针对生成模型合成或编辑的图像篡改检测和定位。

尽管目前的篡改定位方法已取得不小的进展，但在应对多重后处理场景下篡改痕迹被掩盖或破坏的情况时，这些方法在鲁棒性和泛化性能力方面的局限性便凸显出来。各类在线社交网络（OSN, online social network）平台已经成为目前存储和传播图像最便捷的方式之一，且这些平台上传和传播的图像通常经过多种后处理。文献[18-19]指出，几乎所有的 OSN 平台都以有损的方式处理上传的图像。这些有损操作引入的噪声会极大地影响图像篡改模型的定位准确性。以脸书为例，该平台将所有上传的图像转换为像素域，并进行舍入和可能的尺寸调整。随后，应用 JPEG 格式压缩并自适应地选择质量因子（QF）。此外，由于用户的操作，OSN 平台的图像也可能引入各种未知的噪声。这些已知或未知的噪声可能会淹没图像上的篡改痕迹。文献[8, 10, 20]试图通过各种后处理攻击来评估模型在有损图像上的鲁棒性。文献[21]则专注于识别图像操作链中由多个操作按一定顺序组成的操作，采用了基于盲信号分离的特征解耦方法来实现多个操作的识别。对于后处理的伪造图像，SNIS[22]提出了一种

信噪分离的方式，将含有后处理的背景分离出来，弱化后处理操作对篡改检测的负面影响。

然而，上述方法大多只考虑了单一后处理攻击对模型性能的影响，并没有考虑到同时混合多种后处理攻击的情况，忽略了实际应用中可能遇到的多重后处理攻击场景，这限制了它们在真实场景中的适用性。此外，目前大多数模型主要基于卷积神经网络，其卷积操作的固有局限性在于难以捕捉图像的全局相关性特征，从而影响了对图像内容和相互关系的深入理解，这在经过多重后处理攻击的情况下问题尤其严重。

为了克服这些挑战，本文采用了金字塔视觉转换器（PVT）[23]的注意力机制，该机制能在特征提取过程中有效捕捉图像不同区域之间的联系，并通过其金字塔结构在多个尺度上扩展特征的感受野，以最大限度地识别篡改痕迹。同时，采用了UNet[24]架构的变体作为解码器，以优化图像的详细信息和整体结构的处理。在此基础上，本文提出了一种结合了Transformer和CNN的新型图像篡改定位模型，旨在应对多重后处理的挑战。该模型的全局信息捕获能力可以降低社交平台引起的全局失真，从而在篡改定位方面实现了较高的准确性。通过在仿真和实际场景数据集上的实验，证明了所提方法的鲁棒性优于现有的图像篡改定位方法。

本文的主要贡献如下。

1) 相较于传统依赖卷积神经网络的模型，本文采用了金字塔视觉转换器作为编码器的核心网络，其金字塔结构在减少计算量且不增加参数量的前提下，结合注意力机制可以建立全局依赖关系。这使模型能有效捕捉多尺度篡改特征，增强鲁棒性。

2) 在解码器部分，本文对UNet结构进行了改进，其对称的U型结构可以映射高维特征并逐步恢复原始的空间分辨率。这种多尺度特征融合结构能进一步提高模型鲁棒性。

3) 本文使用常见的后处理攻击以及OSN平台[12]进行了广泛的鲁棒性实验。实验结果表明，本文提出的模型在面对经过强后处理的篡改图像时，也能够展现出出色的鲁棒性。

1 模型结构

1.1 模型提出的动机

经OSN平台传输的图像通常会经历多种不同甚至未知的后处理攻击，这些全局失真操作通常削弱了图像篡改特征。因此，要求篡改定位模型具备学习来自多种篡改图像的共同伪造特征的能力，并且能够提取不同语义层面下的篡改特征信息。目前，已有的图像篡改定位模型主要采用卷积操作，但是每个卷积核仅能捕捉输入数据的局部感受野信息，难以提取全局数据之间的长距离特征，这将无法获取丰富且强相关的篡改特征。

Transformer在建模长距离依赖关系方面表现出色，其注意力机制可以有效克服卷积操作所带来的局限性。目前，基于Transformer技术的模型在计算机视觉领域取得了显著的进展。例如，视觉转换器（ViT, vision transformer）[25]在图像分类、识别、分割等任务上展现出比CNN更优的效果。Transformer的注意力机制能够灵活地聚焦在图像的不同区域，从而编码上下文线索，使Transformer具备较强的鲁棒性。文献[26-27]发现注意力机制赋予Transformer很高的网络参数稀疏性，在面对严重遮挡、域转移、空间排列、对抗性、自然扰动等方面展现出较强的鲁棒性。因此，本文选用基于视觉Transformer的金字塔视觉转换器构建编码器，结合金字塔结构以获取多尺度特征。这有助于更好地建立篡改图像中不同区域像素之间的依赖关系，从而提高对篡改痕迹的感知能力。

另一方面，由金字塔视觉转换器得到的多尺度特征需要进行融合和逐层恢复图像分辨率，而类UNet结构的跳跃连接确保了解码过程可以有效地融合不同尺度的特征。同时，类UNet结构通常可以根据任务复杂性扩缩自如，其灵活性使其能够根据特定应用需求来定制模型，而不会过度依赖于特定结构，这使类UNet结构具备一定的鲁棒性[28-30]。目前，计算机视觉领域有很多研究采用类UNet结构进行改进，以提升模型的鲁棒性。

因此，本文设计了金字塔视觉转换器和类UNet结构相结合的模型，以充分提取篡改痕迹。

1.2 模型整体框架

本文提出的模型整体框架如图1所示，整个模型主要由2个部分组成，分别是PVT编码器和类UNet结构的U型解码器。本文的任务是对输入的篡改图像进行定位，找出篡改区域。首先，篡改图像经过PVT编码器模块进行特征提取，随后，得

到的4张特征图经过解码器操作，最终生成预测结果的掩码图。下面将详细介绍PVT编码器和类UNet结构的U型解码器的结构细节。

1.3　PVT编码器

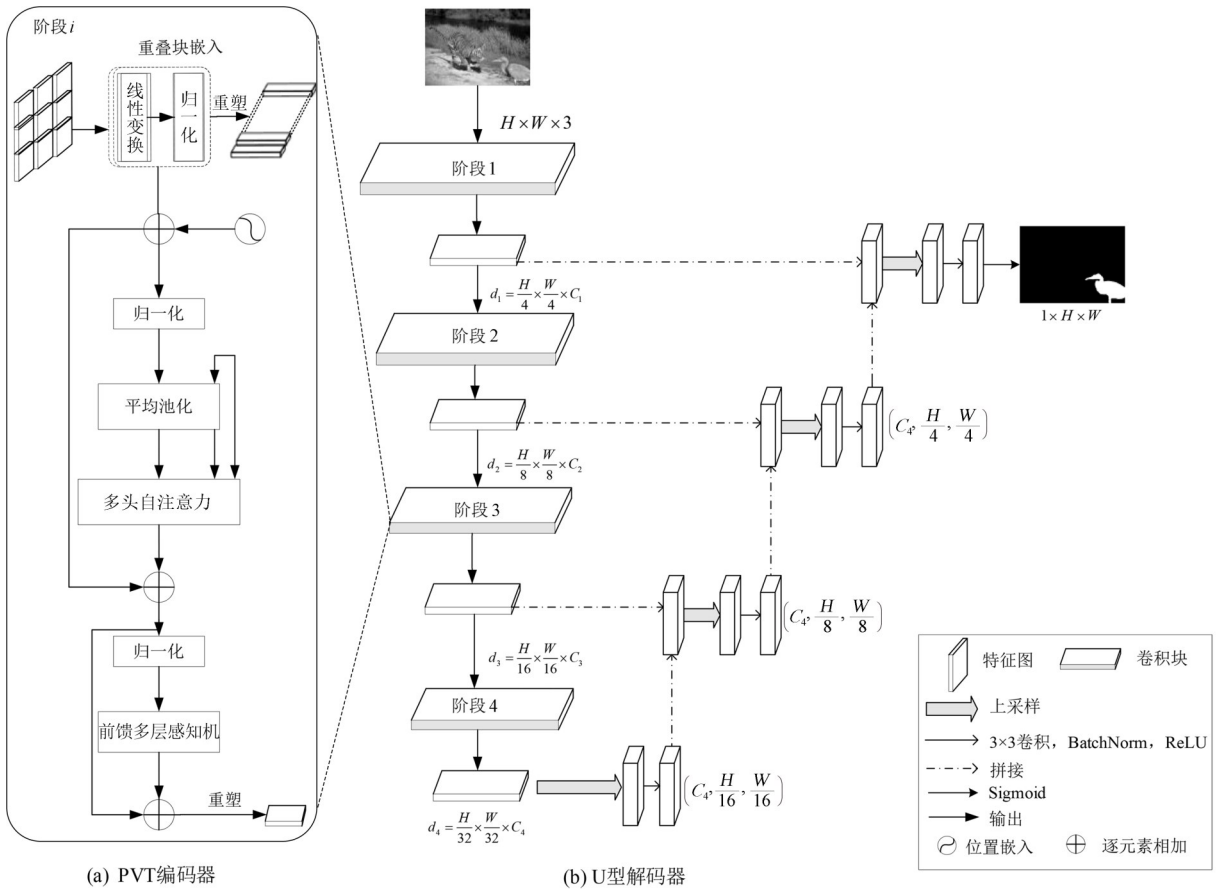
PVT[23]模型是当前表现优异的纯Transformer结构的视觉注意力模型，已被广泛应用于计算机视觉任务。PVT可代替CNN来处理多种下游任务，比如语义分割、目标检测等。

PVT模型是在ViT[25]的基础上改进了多头注意力模块，引入了空间减少注意力（SRA, spatial reduction attention）机制。SRA接收查询Q、键K和值V作为输入，并输出改进后的特征。其中，SRA通过减少K和V的空间维度，有效降低了计算和内存开销。SRA的阶段i的详细信息如式(1)~式(3)所示。

$$\begin{aligned}&\mathrm{SRA}(Q,K,V)=\\&\mathrm{Concat}(\mathrm{head}_0,\mathrm{head}_1,\cdots,\mathrm{head}_j,\mathrm{head}_{N_i})\boldsymbol{W}^o \quad (1)\\&\mathrm{head}_j=\mathrm{Attention}(Q\boldsymbol{W}_j^o,\mathrm{SR}(K)\boldsymbol{W}_j^k,\mathrm{SR}(V)\boldsymbol{W}_j^v) \quad (2)\\&\mathrm{SR}(x)=\mathrm{Norm}(\mathrm{Reshape}(x,R_i)\boldsymbol{W}^s) \quad (3)\end{aligned}$$

其中，$\boldsymbol{W}^o\in R^{C_i\times C_i}$是线性投影矩阵；$\boldsymbol{W}^s\in R^{(R_i^2C_i)\times C_i}$是将输入序列的维数减小到$C_i$的线性投影；$C_i$是阶段$i$的输出通道编号；$N_i$表示阶段$i$ (i = 1,2,3,4)的注意力头数；Attention的计算代价是R^2，R是缩减率；SR是空间缩减率。

本文所采用的编码器是PVT的改进版本PVTv2[31]。相较于PVT，PVTv2引入了零填充卷积来实现重叠块嵌入，从而对图像进行序列化处理。具体地，对于给定尺寸为$H\times W\times 3$的输入，使用步幅大小为S、核尺寸为$2S$–1、填充为S–1的卷积进行处理。同时，PVTv2还采用了线性SRA，通过平均池化将注意力操作前的空间维度$H\times W$缩减为固定大小的$P\times P$，这一策略减少了注意力操作所带来的高计算成本。从图1(a)可以看到，编码器主要由4个相似的阶段组成，堆叠多个独立的Transformer编码器，每个编码器模块由一个多头自注意力（MHSA, multi-head self-attention）机制和一个前馈多层感知机（FMLP, feedforward multi-layer perceptron）组成。前馈多层感知机结构如图2所示，相比PVT，在前馈网络中，第一个全连接层和GeLU激活函数之间添加了3×3深度卷积。



(a) PVT编码器　　(b) U型解码器

图1　模型整体框架

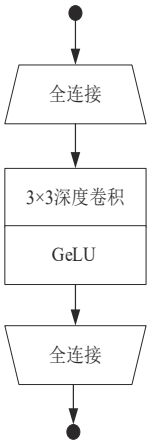


图2 前馈多层感知机结构

输入的图像首先在阶段1经过重叠块嵌入进行划分，得到$\frac{H}{4}\times\frac{W}{4}$个切片，每个切片的尺寸为4 × 4 × 3。随后，将这些切片展平，送入线性投影层，得到$\frac{H}{4}\times\frac{W}{4}\times C_1$的块嵌入。将块嵌入和位置嵌入共同经过第一个Transformer编码器，最终输出尺寸为$\frac{H}{4}\times\frac{W}{4}\times C_1$。同样地，利用前一阶段生成的特征图作为输入，可以获得特征图d_2,d_3,d_4。PVTv2采用了渐进收缩策略，利用块嵌入层来控制特征图的尺度。这里，将阶段i的块大小表示为P_i。在阶段1的开始将输入特征图$d_{i-1}\in R^{H_{i-1}\times W_{i-1}\times C_{i-1}}$平均划分为$\frac{H_{i-1}\times W_{i-1}}{P_i^{\,2}}$个块，然后将每个块展平并投影到$C_i$维嵌入中。经过线性投影后，嵌入块的形状可以视为$\frac{H_{i-1}}{P_i}\times\frac{W_{i-1}}{P_i}\times C_i$，其中高度和宽度是输入的$P_i$倍。这样，在每个阶段就可以灵活地调整特征图的比例尺，形成一个渐进收缩的金字塔结构。

为了让PVTv2适用于图像篡改定位任务，本文移除了最后一个分类层，并在不同阶段生成了4个多尺度特征图，即d_1,d_2,d_3,d_4。d_1提取了图像篡改区域的轮廓信息，d_2,d_3,d_4则提供了更高级的特征。

1.4 U型解码器

本文中的解码器结构采用了类似于UNet的架构，用于解码PVTv2编码器生成的4个不同层级的特征图。整个U型的编码−解码结构中的解码器部分如图1(b)所示。首先通过上采样放大最深层级的特征图d_4，将其尺寸从$\left(C_4,\frac{H}{32},\frac{W}{32}\right)$放大到与较浅层级的特征图$d_3$尺寸相同，即$\left(C_4,\frac{H}{16},\frac{W}{16}\right)$，然后进行拼接降维，得到一个更精细的特征图$f_1$，尺寸为$\left(C_3,\frac{H}{16},\frac{W}{16}\right)$。接着，对特征图$f_1$进行上采样，得到尺寸为$\left(C_3,\frac{H}{8},\frac{W}{8}\right)$的特征图，并与浅层特征图$d_2$拼接，生成尺寸为$\left(C_2,\frac{H}{8},\frac{W}{8}\right)$的特征图$f_2$。同样地，再对特征图$f_2$进行上采样，得到尺寸为$\left(C_2,\frac{H}{4},\frac{W}{4}\right)$的特征图，与浅层特征图$d_1$拼接得到尺寸为$\left(C_1,\frac{H}{4},\frac{W}{4}\right)$的特征图$f_3$。通过多次上采样和拼接操作，逐渐放大和合并特征图，最终得到一个具有高分辨率的特征图f_3。最后，将最终的特征图f_3通过上采样、卷积操作和Sigmoid函数激活，生成最终的篡改区域定位掩码图。解码结构中的上采样操作都是使用转置卷积，也称为反卷积，是一种卷积神经网络中常用的操作，通常用于图像处理和深度学习中的上采样和特征映射还原。与常见的双线性插值、最近邻插值不同，转置卷积使用带有学习参数的卷积核进行运算，更加拟合深度学习模型参数优化的方式。这一点在消融实验上得到了很好的验证。

本文使用的损失函数是在图像篡改定位模型中常用的二分类交叉熵损失函数BCELoss，如式(4)所示。

$$\text{BCELoss}=-\frac{1}{n}\sum(y\log(p)+(1-y)\log(1-p)) \quad (4)$$

BCELoss的实现简单且计算效率高，在本文实验中表现出不错的性能。

2 实验结果

本节先通过多种不同的数据后处理方式，如随机裁剪、JPEG压缩、高斯模糊和高斯噪声，对所提模型的鲁棒性进行了分析和比较。同时，在OSN平台上与主流模型进行了鲁棒性比较。最后，评估了所提模型在多个公开数据集上的性能，并与当前主流的图像篡改定位模型进行了比较，以证明本文方法的有效性。

2.1 实验数据集

本文实验使用如下公开的数据集进行实验。

Columbia[32]。该数据集包含183张真实图像和

180张经过简单拼接的篡改图像。

NIST16[33]。该数据集包含564张篡改图像，主要应用了拼接、复制移动和删除3种篡改方式。

CASIA[34]。该数据集包括CASIAv1和CASIAv2这2个部分。CASIAv1涵盖960张篡改图像，CASIAv2包含5 123张篡改图像。这些图像的篡改方式包括拼接、复制移动，同时还应用了裁剪、旋转、模糊等后处理操作。

IMD2020[35]。该数据集的篡改图像从互联网收集而来，共有2 010张篡改图像，涵盖了拼接、复制移动、删除等篡改方式。

Coverage[36]。该数据集是一个专为复制移动设计的较小数据集，包含100张篡改图像。

DSO[37]。该数据集包含100张拼接的篡改人脸图像。

OSN[12]。该数据集包括来自CASIAv1数据集的920张、Columbia数据集的160张、NIST16数据集的564张以及DSO数据集的100张篡改图像。这些篡改图像都被上传到社交媒体平台，如微信、脸书、微博和WhatsApp，以获得对应的社交媒体平台后处理版本。

本文的训练方式分为3种：基准训练、预训练和微调。基准训练模式下，模型在特定数据集上训练并测试。预训练模式下，模型在外部数据集上训练后直接在目标数据集上测试，展示模型对不同数据分布的适应性。微调模式结合了预训练权重和目标数据集的训练，进一步提升模型的针对性和准确性。针对基准训练，本文根据SAT-IFL[10]中的训练−测试比例，拆分配置CASIA、Columbia、NIST16、IMD2020和Coverage这5个数据集，以进行模型的重新训练和评估。其中训练集占75%，测试集占25%。表1展示了各数据集基准训练时包含的样本数量。而预训练方面，本文在CASIAv2数据集上训练得到预训练模型，并在OSN的DSO、CASIAv1和NIST16这3个数据集上进行鲁棒性测试。为了比较模型的篡改定位性能，本文考虑了6种主流的方法，分别是RGB-N[6]、DenseFCN[7]、ManTraNet[8]、MVSSNet[9]、SAT-IFL[10]和IF-OSD[12]，并在微调模式下比较了RGB-N和本文模型。此外，本文在所有的训练实验中默认添加旋转和翻转2种数据增强方式。在本文的实验中，输入图像经过预处理后的尺寸大小为3 × 512 × 512。通过PVTv2的4个阶段，得到的特征图通道数和尺寸分别是（64, 128, 128）、（128, 64, 64）、（320, 32, 32），以及（512, 16, 16）。所有的实验均在PyTorch框架中实现，并在NVIDIA A100-PICE-40GB上进行训练。训练批次大小设定为32，进行100轮训练，选择效果最好的一轮。初始的学习率设置为1.0×10^{-4}，优化器采用自适应矩估计（Adam, adaptive moment estimation）优化器。

表 1　各数据集基准训练时包含的样本数量

数据集	篡改类型	训练−测试分割		总数
		训练	测试	
Columbia	拼接、复制移动	135	45	180
NIST16	多种类型	423	141	564
CASIA	拼接、复制移动	5 123	920	6 043
IMD2020	多种类型	1 507	503	2 010
Coverage	复制移动	75	25	100

2.2　评价指标

根据之前的工作[7-10]，本文将像素级F1分数、交并比（IoU, intersection over union）和接受者操作特征（ROC）曲线下面积（AUC, area under the curve）这3个指标作为模型定位性能比较的评估指标。

F1分数是精度（Precision）和召回率（Recall）的调和均值。IoU在篡改定位中被定义为模型预测的篡改区域和真实篡改掩码之间的重叠部分占两者并集的比例。AUC在篡改定位中则是以像素预测结果为单位，用于衡量二分类模型的性能。当AUC=0.5时，表示模型的预测与随机模型相当，具体计算式为

$$\mathrm{F1} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{5}$$

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{6}$$

$$\mathrm{AUC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{7}$$

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。F1和IoU指标综合考虑了查全率和查准率，相比AUC更能反映图像篡改定位模型的性能表现。

2.3　鲁棒性实验

在实际场景中，篡改图像往往会经过各种各样以及不同程度的后处理。为了评估所提模型的鲁棒性，本文采用基准训练模式，并在IMD2020和NIST16数据集上对该模型在多种不同强度的后处

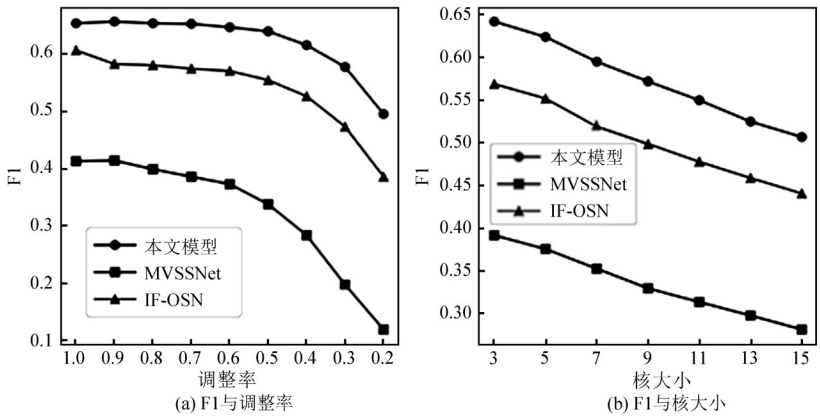
理攻击下的性能进行了评估。这些后处理攻击包括缩放（调整率为0.25×和0.78×）、高斯模糊（核参数为3和15）、高斯噪声（标准差为3和15）以及JPEG压缩（质量因子为50、60、80和100）。为确保公平，所有模型按照相同的实验设置进行训练和测试。将本文模型与MVSSNet[9]和IF-OSN[12]进行了比较，实验结果如表2所示，其中，加粗字体表示最高值。数据显示，在IMD2020数据集上，本文模型在面对各种不同强度的后处理攻击时明显优于其他2个模型，F1、IoU和AUC指标均表现出更好的结果，尤其是在高斯噪声（15）的攻击下，本文模型的F1指标下降幅度为57%，而MVSSNet和IF-OSN分别下降了100%和87%。这表明本文模型在高斯噪声方面具有更强的抗干扰性；在NIST16数据集上，本文模型的F1、AUC和IoU指标明显优于MVSSNet模型。相比IF-OSN，本文模型初始的F1和AUC指标要低一些，但是在面对高斯噪声、高斯模糊、JPEG压缩（50）、JPEG压缩（60）以及缩放（0.25×）攻击时，本文模型表现更出色。在JPEG压缩（80）、JPEG压缩（100）以及缩放（0.78×）攻击下，本文模型和IF-OSN的性能基本保持不变。从图3可以更直观地观察到本文模型在鲁棒性方面的优势，在IMD2020数据集上，本文模型在F1指标上明显优于其他2个模型。从图4可以看到，本文模型在NIST16数据集上相对于MVSSNet表现出更明显的优势，而与IF-OSN相比，本文模型在面对不同方式、不同强度的后处理攻击时下降趋势更平缓，进一步说明本文模型在篡改定位方面的可靠性。

2.4 OSN实验

OSN提供了全新的挑战情景，在这些场景中，图像经常被压缩和调整，可能还包含未知的噪声，从而导致图像中篡改痕迹被减弱。2.3节的鲁棒性实验中定量分析了不同的单一后处理攻击对模型性能的影响。与单一后处理不同，OSN的处理操作更多样和不可预测。因此，本文在OSN上评估了模型的鲁棒性，并与DenseFCN[7]、MVSSNet[9]和IF-OSN[12]进行了对比。本文选取了在CASIAv2数据集上F1、IoU和AUC表现最好的预训练模型，将这些模型在OSN平台上进行评估。为了展示本文数据增强方式的有效性，本文也将经过数据增强之后的预训练模型测试结果展示在表3中，其中加粗字体表示每一列的最高值。从表3中可以观察到，无论是通过微信、微博、脸书还是WhatsApp，在OSN的NIST16和CASIAvl数据集上，本文模型在F1、AUC和IoU都表现最佳。而在DSO数据集上，在训练过程中没有经过数据增强的模型表现要稍微优于在训练过程中加入数据增强的模型。其原因可能是DSO数据集相对较小，应用数据增强后的模型可能过于关注增强后的样本，从而忽略了原始数据的信息。通过平均F1、AUC和IoU值的综合考量，本文模型无论是在数据增强前还是在增强后，性能都优于其他模型。通过在OSN平台上的鲁棒性实验可以发现，所有的模型在经历社交平台上传之后性能都普遍下降，这可能是由于社交平台上传过程中引入的未知变化或噪声。本文模型相比其他模型在抵抗这些真实场景的篡改定位时效果稳

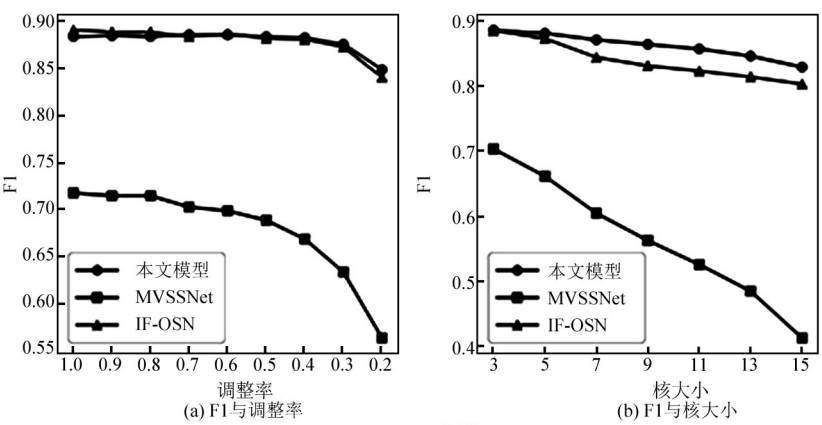
表 2　本文模型与MVSSNet、IF-OSN在IMD2020、NIST16数据集上的常见后处理攻击鲁棒性分析

操作	IMD2020			NIST16		
	本文模型	MVSSNet	IF-OSN	本文模型	MVSSNet	IF-OSN
	F1, AUC, IoU	F1, AUC, IoU	F1, AUC, IoU	F1, AUC, IoU	F1, AUC, IoU	F1, AUC, IoU
对照（无数据增强）	**0.65, 0.95, 0.57**	0.41 ,0.84, 0.32	0.61 ,0.94, 0.52	0.88 ,**0.99,** 0.83	0.72 ,0.93, 0.63	**0.89 ,0.99, 0.84**
缩放（0.78×）	**0.65, 0.95, 0.57**	0.40, 0.83, 0.30	0.58, 0.93, 0.50	0.88, **0.99,** 0.83	0.71, 0.93, 0.62	**0.89,** 0.98, **0.84**
缩放（0.25×）	**0.54, 0.91, 0.46**	0.17, 0.75, 0.12	0.44, 0.87, 0.36	**0.87, 0.99, 0.82**	0.61, 0.92, 0.52	0.85, 0.98, 0.80
高斯模糊（3）	**0.65, 0.95, 0.56**	0.39, 0.84, 0.29	0.57, 0.93, 0.48	**0.89, 0.99, 0.83**	0.70, 0.93, 0.61	**0.89,** 0.98, **0.83**
高斯模糊（15）	**0.54, 0.91, 0.46**	0.28, 0.79, 0.20	0.44, 0.86, 0.35	**0.83, 0.99, 0.77**	0.41, 0.86, 0.33	0.80, 0.97, **0.77**
高斯噪声（3）	**0.59, 0.93, 0.45**	0.30, 0.80, 0.32	0.50, 0.92, 0.41	**0.80, 0.97, 0.73**	0.57, 0.88, 0.47	**0.80, 0.97, 0.73**
高斯噪声（15）	**0.28, 0.80, 0.22**	0.00, 0.53, 0.00	0.10, 0.67, 0.06	**0.29, 0.79, 0.22**	0.14, 0.69, 0.09	0.19, 0.71, 0.14
JPEG压缩（100）	**0.65, 0.94, 0.56**	0.41, 0.84, 0.31	0.60, 0.94, 0.52	0.88, **0.99**, 0.81	0.72, 0.93, 0.63	**0.89, 0.99, 0.84**
JPEG压缩（80）	**0.61, 0.93, 0.54**	0.39, 0.83, 0.30	0.55, 0.93, 0.47	0.88, **0.99**, 0.82	0.72, 0.93, 0.63	**0.89, 0.99, 0.84**
JPEG压缩（60）	**0.59, 0.93, 0.51**	0.36, 0.83, 0.28	0.52, 0.92, 0.41	**0.87, 0.99, 0.82**	0.72, 0.92, 0.63	**0.87,** 0.98, **0.82**
JPEG压缩（50）	**0.58, 0.93, 0.49**	0.35, 0.82, 0.26	0.52, 0.92, 0.44	**0.88, 0.99, 0.82**	0.72, 0.93, 0.63	0.87, 0.98, **0.82**



(a) F1与调整率　(b) F1与核大小　(c) F1与标准差　(d) F1与质量因子

图3　本文模型与MVSSNet、IF-OSN在IMD2020上的F1分析



(a) F1与调整率　(b) F1与核大小　(c) F1与标准差　(d) F1与质量因子

图4　本文模型与MVSSNet、IF-OSN在NIST16上的F1分析

表 3 本文模型与 DenseFCN、MVSSNet、IF-OSN 在 CASIAv2 的预训练模型在 OSN 平台上的鲁棒性分析

模型	OSN	NIST16			DSO			CASIAv1			Coverage		
		F1	AUC	IoU	F1	AUC	IoU	F1	AUC	IoU	F1	AUC	IoU
DenseFCN	None	0.12	0.59	0.08	0.24	0.60	0.14	0.15	0.55	0.09	0.17	0.58	0.10
MVSSNet	None	0.24	0.72	0.17	0.07	0.65	0.04	0.45	0.81	0.35	0.25	0.73	0.19
IF-OSN	None	0.29	0.78	0.25	0.24	0.73	0.18	0.52	0.86	0.47	0.35	0.79	0.30
本文模型	None	0.35	0.65	0.30	**0.35**	0.64	**0.28**	0.58	0.79	0.52	0.43	0.69	0.37
本文模型（数据增强）	None	**0.36**	**0.83**	**0.31**	0.22	**0.76**	0.18	**0.70**	**0.93**	**0.65**	**0.43**	**0.84**	**0.38**
DenseFCN	脸书	0.12	0.59	0.08	0.24	0.61	0.14	0.15	0.55	0.09	0.17	0.58	0.10
MVSSNet	脸书	0.23	0.72	0.17	0.07	0.65	0.04	0.42	0.81	0.33	0.24	0.73	0.18
IF-OSN	脸书	0.28	0.77	0.23	0.27	0.76	0.21	0.46	0.84	0.40	0.33	0.79	0.28
本文模型	脸书	0.35	0.66	0.30	**0.36**	0.65	**0.29**	0.52	0.76	0.45	0.41	0.69	0.35
本文模型（数据增强）	脸书	**0.36**	**0.83**	**0.31**	0.22	**0.76**	0.18	**0.67**	**0.93**	**0.61**	**0.42**	**0.84**	**0.37**
DenseFCN	微博	0.12	0.59	0.08	0.24	0.61	0.14	0.15	0.55	0.09	0.17	0.58	0.10
MVSSNet	微博	0.25	0.71	0.18	0.08	0.67	0.05	0.41	0.80	032	0.25	0.73	0.11
IF-OSN	微博	0.25	0.77	0.21	0.21	0.73	0.15	0.44	0.84	0.40	0.30	0.78	0.25
本文模型	微博	0.25	0.5	0.21	**0.35**	**0.64**	**0.28**	0.51	0.76	0.46	0.37	0.66	0.32
本文模型（数据增强）	微博	**0.35**	**0.83**	**0.31**	0.21	0.75	0.17	**0.63**	**0.92**	**0.58**	**0.40**	**0.83**	**0.35**
DenseFCN	微信	0.12	0.59	0.08	0.24	0.61	0.14	0.15	0.55	0.09	0.17	0.59	0.10
MVSSNet	微信	0.24	0.71	0.18	0.09	0.65	0.05	0.40	0.80	0.30	0.24	0.72	0.18
IF-OSN	微信	0.27	0.76	0.22	0.23	**0.75**	0.17	0.36	0.81	0.30	0.29	0.77	0.23
本文模型	微信	0.34	0.67	0.29	**0.34**	0.64	**0.27**	0.44	0.71	0.37	**0.37**	0.67	0.31
本文模型（数据增强）	微信	**0.35**	**0.81**	**0.30**	0.20	0.74	0.16	**0.57**	**0.91**	**0.51**	0.37	**0.82**	**0.32**
DenseFCN	WhatsApp	0.12	0.59	0.08	0.24	0.61	0.14	0.15	0.55	0.09	0.17	0.59	0.10
MVSSNet	WhatsApp	0.25	0.72	0.18	0.09	0.66	0.06	0.42	0.81	0.32	0.25	0.73	0.19
IF-OSN	WhatsApp	0.28	0.79	0.24	0.25	0.75	0.19	0.49	0.86	0.43	0.34	0.80	0.29
本文模型	WhatsApp	**0.35**	0.66	0.30	**0.35**	0.64	**0.28**	0.52	0.76	0.45	**0.41**	0.69	0.35
本文模型（数据增强）	WhatsApp	0.35	**0.83**	**0.30**	0.20	**0.75**	0.16	**0.67**	**0.93**	**0.61**	0.40	**0.84**	**0.36**

定，这一结果表明，本文模型具备不错的泛化能力和鲁棒性。图 5 的可视化结果可以更直观地体现本文模型的鲁棒性。在受到社交平台上各种不同的强后处理操作影响的情况下，本文模型的性能仍能保持稳定。其中，DenseFCN 模型的可视化图呈现大片白色，只有极少数的结果有零星的黑点。从表 3 的 DenseFCN 模型在 CASIAv1 的实验结果来看，DenseFCN 在图像篡改定位的准确性和其他性能指标方面都表现较差，这直接影响了其可视化效果的质量。本文将模型得到的定位结果经过阈值化后得到 0 和 1 的二值化图。随机选择一张未阈值化的结果图进行分析，发现其预测值在 0.500 0 至 0.500 4 之间。这些值阈值化后变为 1，从而在可视化图中呈现大量白色。这表明 DenseFCN 模型存在较高的虚警率，错误地将非篡改区域识别为篡改区域，反映了该模型在性能上的不足。

2.5 无后处理场景下的模型性能比较

2.3 节和 2.4 节测试了本文模型在有后处理场景下的鲁棒性，本节在无后处理场景下的原始数据集上对比模型的定位性能。从表 4 的结果可以看出，本文模型在 5 个基准数据集上的平均 F1 和 AUC 性能指标均优于 RGB-N[6]、DenseFCN[7]、ManTraNet[8]、MVSSNet[9]、SAT-IFL[10] 以及 IF-OSN[12]。尤其是在微调模式下，本文模型的性能超越了基于 COCO[38] 数据集预训练的 RGB-N 模型。这表明即使在小型数据集 CASIAv2 上进行预训练，本文模型也能展现出良好的效果。在基准训练模式下，与 MVSSNet 相比，本文模型的 F1 指标平均提升了 0.141。而与 IF-OSN 模型相比，本文模型仅在 Coverage 数据集稍差一些。通过对比可以观察到，本文模型经过数据增强之后在数据集 CASIAv1 和 IMD2020 的 F1 提升效果最显著，分别比 DenseFCN 高 0.55 和 0.53，在 Columbia 数据集上 F1 达到了 0.99。图 6 展示了本文模型与 MVSSNet、DenseFCN 和 IF-OSN 在 Columbia、NIST16、CASIAv1、IMD2020 以及 Coverage 这 5 个公开数据集上的可视化结果。从图 6 中可以直观地观察到，DenseFCN 的预测效果并不理想，虚警率过高，表明这个模型性能上的不足。MVSSNet 和 IF-OSN 也存在虚警率较高的问题，尽管能够大致确定篡改区域，但是依然存在细节不够精确的问题，并不能准确地定位出篡改区域。而本文模型预测效果与真实掩码非常相近，能够准确地定位出图像篡改区域的位置，充分证明了模型优秀的定位能力。

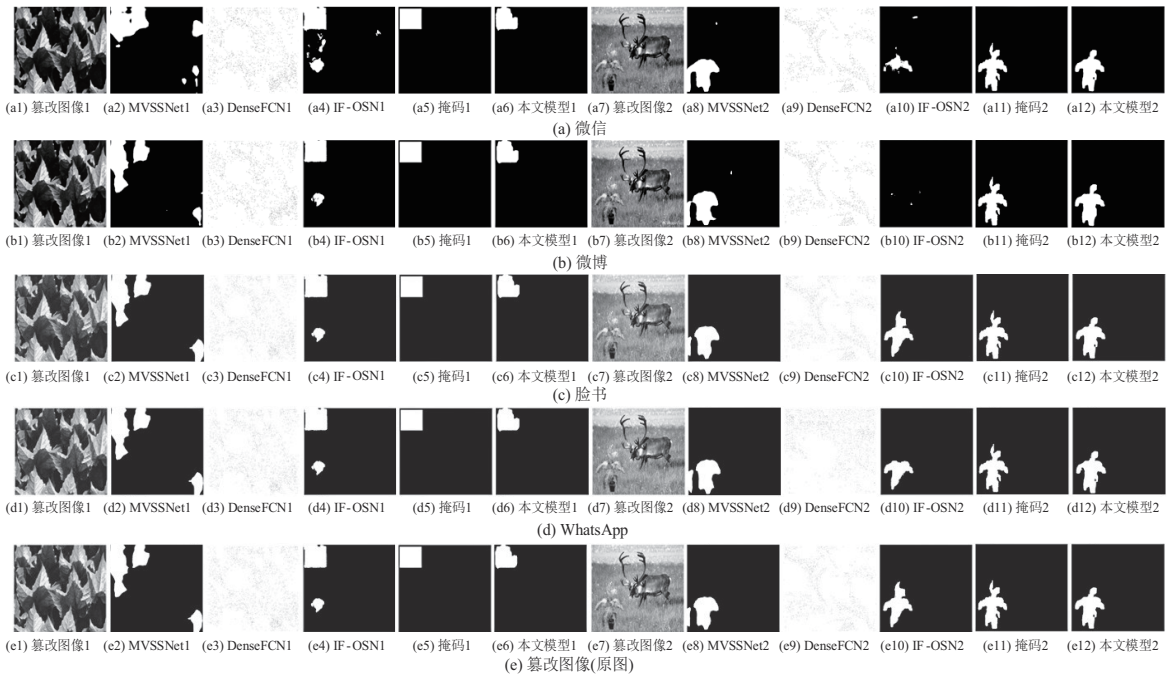


图5　本文模型与MVSSNet、DenseFCN和IF-OSN在OSN平台上的CASIAv1数据集的可视化结果

表4　本文所提模型及6种对比模型在Columbia、CASIA、NIST16、IMD2020、Coverage上的定位结果

模型	训练模式	Columbia		CASIAv1		NIST16		IMD2020		Coverage		平均	
		AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
RGB-N	微调	0.86	0.70	0.80	0.41	0.94	0.72	—	—	0.82	0.44	0.85	0.57
ManTraNet	基准训练	0.99	0.98	0.65	0.09	0.71	0.14	0.79	0.07	0.95	0.76	0.82	0.41
SAT-IFL	基准训练	0.92	0.89	0.79	0.38	0.94	0.62	—	—	0.86	0.53	0.88	0.61
DenseFCN	基准训练	0.90	0.41	0.55	0.15	0.68	0.12	0.63	0.13	0.59	0.20	0.67	0.20
MVSSNet	基准训练	0.99	0.92	0.81	0.45	0.93	0.72	0.82	0.42	0.90	0.55	0.89	0.61
IF-OSN	基准训练	0.97	0.87	0.86	0.52	0.99	0.89	0.94	0.61	0.97	0.77	0.95	0.73

2.6　消融实验

为了展示本文对模型所进行改进的有效性，本文进行了消融研究，以评估在不同设置下逐步添加组件对模型的影响。首先，尝试了多种解码器的模型网络结构，包括空洞卷积、转置卷积，以及两者的结合。对不同的损失函数、初始学习率、优化器和批次大小进行了对比实验，分析这些条件对模型性能的影响。最后，进行了数据增强的消融实验。以上实验统一在数据集CASIAv2上训练，CASIAv1上测试得到相关数据进行比较。实验设置损失函数为BCELoss，初始学习率为0.000 1，选择Adam优化器以及批次大小为32作为对照。在CASIAv1上的定量结果如表5~表10所示，加粗字体表示最高值。

1) 空洞卷积和转置卷积

转置卷积和空洞卷积的消融实验结果如表5所示。空洞卷积通过控制间隔大小扩展感受野，同时保持特征图尺寸，从而丰富空间信息，扩展模型视野。本文模型在未应用空洞卷积前，初始F1性能为0.45；应用后，F1提升至0.52，显示出显著效果。本文还采用转置卷积替代传统上采样，以避免信息丢失。转置卷积通过卷积核尺寸、步长设置精细调控输出图像尺寸和分辨率。在本文模型中，上采样层的核尺寸为2，步幅为2，填充为0；最后一层核尺寸为8，步幅为4，填充为2。实验表明，转置卷积将CASIAv1上的F1指标从0.45提升至0.58，增加了10%。空洞卷积和转置卷积结合使用可实现上采样和细化，但根据实验，两者结合并非总是优于单独使用。这可能因为转置卷积引入上采样步骤导致重叠和混叠，而空洞卷积旨在扩大感受野，两者目标不完全一致。

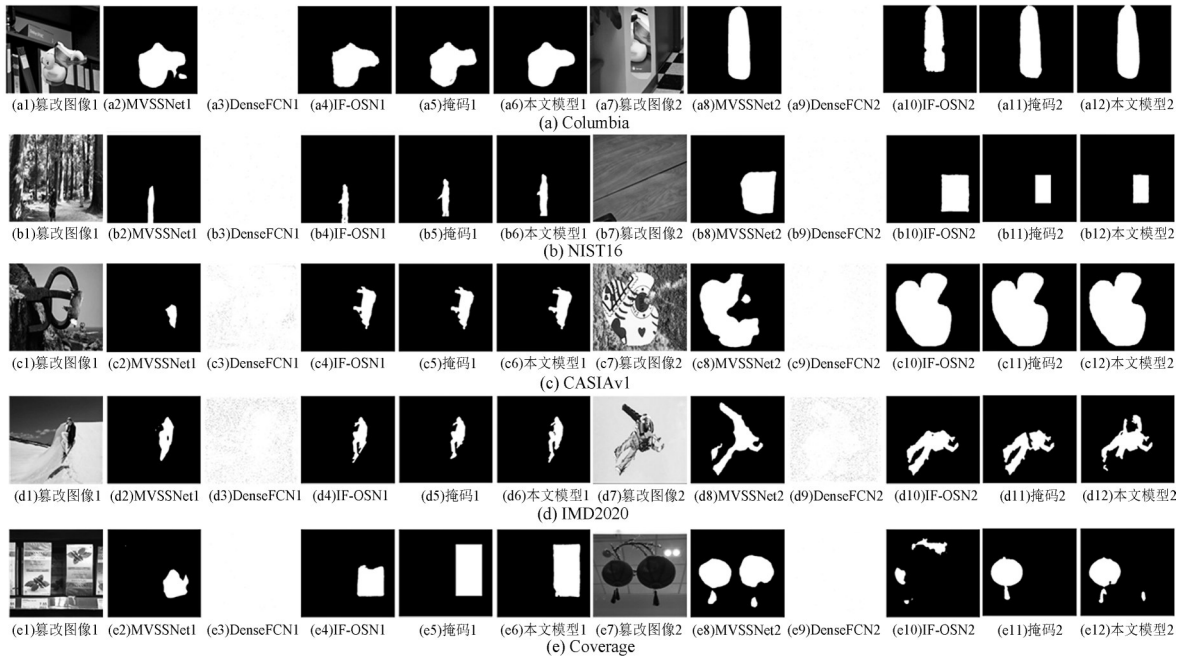


图 6　本文模型与 DenseFCN、MVSSNet 和 IF-OSN 在 5 个数据集上的可视化结果

表 5　　转置卷积和空洞卷积的消融实验结果

实验设置	F1	AUC	IoU
对照	0.45	0.82	0.41
空洞卷积	0.52	0.82	0.47
转置卷积	**0.58**	0.79	**0.52**
转置卷积+空洞卷积	0.54	**0.83**	0.50

2) 损失函数

本文比较了 BCELoss、DiceLoss 和 BCEDiceLoss 这 3 种常用的损失函数。BCELoss 通过测量模型输出与实际标签之间的二分类交叉熵来评估模型的性能。DiceLoss 评估模型预测的二进制掩膜与实际二进制掩膜的相似度，其值越小表示相似度越高。BCEDiceLoss 是两者的结合。相比其他 2 个损失函数，BCELoss 对像素级别的错误更敏感，配合本文模型的训练可以达到不错的训练效果。不同损失函数的影响如表 6 所示，相比次优的 DiceLoss，BCELoss 的 F1 的指标高 0.02。

3) 初始学习率

为了研究不同的初始学习率对模型训练的影响，本文在优化的初始学习率 0.000 1 附近，设置了 0.000 2 和 0.000 05 的对比实验，配合 Adam 优化器。从表 7 的结果可以看出，设置更小的学习率 0.000 1 和 0.000 05 更适合本文模型。这说明本文模型具有大量的参数和层次，较小的初始学习率有助于实现良好的训练效果。过大的学习率可能导致参数更新过快，无法收敛到合适的解。考虑到更小的初始学习率需要花费的时间更长，所以本文统一选取初始学习率为 0.000 1 进行实验。

表 6　　不同损失函数的影响

损失函数	F1	AUC	IoU
BCELoss	**0.58**	0.79	**0.52**
DiceLoss	0.56	0.76	0.49
BCEDiceLoss	0.54	**0.85**	0.48

表 7　　不同的初始学习率的影响

学习率	F1	AUC	IoU
0.000 2	0.48	0.85	0.43
0.000 1	0.58	0.79	0.52
0.000 05	**0.63**	**0.91**	**0.57**

4) 优化器

本文对 2 种主要的优化器进行了对比研究，具体选择了随机梯度下降（SGD, stochastic gradient descent）和 Adam 进行实验。SGD 通过计算每个样本或小批量样本的梯度，并利用学习率更新模型权重。相比之下，Adam 结合了动量和自适应学习率的思想，通过梯度的指数移动平均计算动量，并适应每个参数的学习率。从表 8 的对比结果来看，Adam 优化器表现较佳，因此本文选择 Adam 优化器进行实验。

表 8　不同优化器的影响

优化器	F1	AUC	IoU
Adam	**0.58**	**0.79**	**0.52**
SGD	0.24	0.76	0.17

5) 批次大小

本文比较了批次大小为 8、16 和 32 的效果。由于计算资源限制，批次大小最大为 32。表 9 的结果显示，在充分利用计算资源和加快训练速度的前提下，批次大小为 32 的效果相比其他的好一些，但与 16 的差距不大。在计算资源有限时，通常也会选择较少批次来实验，配合合适的优化器和学习率调整器也可以达到不错的效果。

表 9　不同批次大小的影响

批次大小	F1	AUC	IoU
8	0.46	0.84	0.42
16	0.56	**0.85**	0.46
32	**0.58**	0.79	**0.52**

6) 数据增强

此外，本文在训练过程中增加了数据增强的方式。具体表现为每一轮训练随机加入至少 1 种、至多 7 种的数据增强方式，分别是 JPEG 压缩、高斯噪声扰动、高斯模糊、随机裁剪、旋转、水平翻转和垂直翻转。这些增强方式的参数随机设置，可以最大程度模拟现实场景。在训练中，随机应用数据增强的比例为 0.2，经过实验验证，该比例取得的性能最优。从表 10 的结果可以看到，本文模型在经过数据增强之后 F1 从 0.58 提升到 0.70，性能提升了 13.6%。

表 10　不同随机比例的数据增强影响

随机比例 n	F1	AUC	IoU
n=0.1	0.67	0.91	0.62
n=0.2	**0.70**	**0.93**	**0.65**
n=0.3	0.70	0.93	0.64
n=0.5	0.68	0.92	0.62

3　结束语

针对各社交平台对图像的有损操作，本文提出了一个基于 PVT 和类 UNet 架构的图像篡改定位模型，该模型综合了 PVT 在特征表示方面的强大能力和多尺度信息，结合 U 型解码器，能够恢复图像的细节和空间分辨率，从而有效地应对图像篡改定位的挑战。实验结果表明，所提模型在多个公开的数据集上的性能优于其他主流的图像篡改定位模型。社交平台上的实验表明，即使经过不同强度的后处理，本文模型依然具备出色的性能。此外，鲁棒性实验验证了本文模型在高斯模糊、高斯噪声、JPEG 压缩和调整大小的后处理攻击时的有效性和鲁棒性。

提高图像篡改定位模型的鲁棒性是实际应用场景中一个很重要的问题。在未来工作中，笔者将进一步探索模型的扩展和改进，特别是多模态数据融合方面。计划引入文本描述、图像的元数据（包括拍摄设备、拍摄时间等）以及图像的上下文信息等不同类型数据。这些多维度信息的结合将有助于模型更全面地理解图像内容和语境，从而增强其在复杂应用场景下的适应性和鲁棒性。

参考文献：

[1] 李晓龙, 俞能海, 张新鹏, 等. 数字媒体取证技术综述[J]. 中国图象图形学报, 2021, 26(6): 1216-1226.
LI X L, YU N H, ZHANG X P, et al. Overview of digital media forensics technology[J]. Journal of Image and Graphics, 2021, 26(6): 1216-1226.

[2] 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.

[3] CHIERCHIA G, POGGI G, SANSONE C, et al. A Bayesian-MRF approach for PRNU-based image forgery detection[J]. IEEE Transactions on Information Forensics and Security, 2014, 9(4): 554-567.

[4] KORUS P, HUANG J W. Multi-scale fusion for improved localization of malicious tampering in digital images[J]. IEEE Transactions on Image Processing, 2016, 25(3): 1312-1326.

[5] 李昊东, 庄培裕, 李斌. 基于深度学习的数字图像篡改定位方法综述[J]. 信号处理, 2021, 37(12): 2278-2301.
LI H D, ZHUANG P Y, LI B. A survey on deep learning based digital image tampering localization methods[J]. Journal of Signal Processing, 2021, 37(12): 2278-2301.

[6] ZHOU P, HAN X T, MORARIU V I, et al. Learning rich features for image manipulation detection[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 1053-1061.

[7] ZHUANG P Y, LI H D, TAN S Q, et al. Image tampering localization

using a dense fully convolutional network[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 2986-2999.

[8] 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 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 9535-9544.

[9] CHEN X R, DONG C B, JI J Q, et al. Image manipulation detection by multi-view multi-scale supervision[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 14165-14173.

[10] ZHUO L, TAN S Q, LI B, et al. Self-adversarial training incorporating forgery attention for image forgery localization[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 819-834.

[11] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv Preprint, arXiv: 1412.6572, 2014.

[12] WU H W, ZHOU J T, TIAN J Y, et al. Robust image forgery detection against transmission over online social networks[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 443-456.

[13] GUO X, LIU X H, REN Z Y, et al. Hierarchical fine-grained image forgery detection and localization[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 3155-3165.

[14] BI X L, YAN W Q, LIU B, et al. Self-supervised image local forgery detection by JPEG compression trace[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(1): 232-240.

[15] WU H W, ZHOU J T, ZHANG S L. Generalizable synthetic image detection via language-guided contrastive learning[J]. arXiv Preprint, arXiv: 2305.13800, 2023.

[16] LORENZ P, DURALL R L, KEUPER J. Detecting images generated by deep diffusion models using their local intrinsic dimensionality[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Piscataway: IEEE Press, 2023: 448-459.

[17] TANTARU D, ONEATA E, ONEATA D. Weakly-supervised deepfake localization in diffusion-generated images[J]. arXiv Preprint, arXiv: 2311.04584, 2023.

[18] SUN W W, ZHOU J T, LI Y M, et al. Robust high-capacity watermarking over online social network shared images[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(3): 1208-1221.

[19] SUN W W, ZHOU J T, LYU R, et al. Processing-aware privacy-preserving photo sharing over online social networks[C]//Proceedings of the 24th ACM international conference on Multimedia. New York: ACM Press, 2016: 581-585.

[20] LIU X H, LIU Y J, CHEN J, et al. PSCC-Net: progressive spatio-channel correlation network for image manipulation detection and localization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(11): 7505-7517.

[21] CHEN J X, LIAO X, WANG W, et al. A features decoupling method for multiple manipulations identification in image operation chains[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2021: 2505-2509.

[22] CHEN J X, LIAO X, WANG W, et al. SNIS: a signal noise separation-based network for post-processed image forgery detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(2): 935-951.

[23] WANG W H, XIE E Z, LI X, et al. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 548-558.

[24] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2015: 234-241.

[25] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale[J]. arXiv Preprint, arXiv: 2010.11929, 2020.

[26] LIU A S, TANG S Y, LIANG S Y, et al. Exploring the relationship between architecture and adversarially robust generalization[J]. arXiv Preprint, arXiv: arXiv: 2209.14105, 2022.

[27] NASEER M, RANASINGHE K, KHAN S, et al. Intriguing properties of vision transformers[J]. arXiv Preprint, arXiv: 2105.10497, 2021.

[28] ZHOU Y M, YING Q C, WANG Y F, et al. Robust watermarking for video forgery detection with improved imperceptibility and robustness[C]//Proceedings of the 2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP). Piscataway: IEEE Press, 2022: 1-6.

[29] HUANG Y H, BIAN S, LI H D, et al. DS-UNet: a dual streams UNet for refined image forgery localization[J]. Information Sciences, 2022, 610: 73-89.

[30] LIU B, WU R L, BI X L, et al. D-UNet: a dual-encoder U-Net for image splicing forgery detection and localization[J]. arXiv Preprint, arXiv: 2012.01821, 2020.

[31] WANG W H, XIE E Z, LI X, et al. PVT v2: improved baselines with pyramid vision transformer[J]. Computational Visual Media, 2022, 8(3):

415-424.

[32] NG T T, CHANG S F. A data set of authentic and spliced image blocks[R]. Columbia University, Advent Technical Report, 2004.

[33] GUAN H Y, KOZAK M, ROBERTSON E, et al. MFC datasets: large-scale benchmark datasets for media forensic challenge evaluation[C]//Proceedings of the 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). Piscataway: IEEE Press, 2019: 63-72.

[34] DONG J, WANG W, TAN T N. CASIA image tampering detection evaluation database[C]//Proceedings of the 2013 IEEE China Summit and International Conference on Signal and Information Processing. Piscataway: IEEE Press, 2013: 422-426.

[35] NOVOZÁMSKÝ A, MAHDIAN B, SAIC S. IMD2020: a large-scale annotated dataset tailored for detecting manipulated images[C]//Proceedings of the 2020 IEEE Winter Applications of Computer Vision Workshops (WACVW). Piscataway: IEEE Press, 2020: 71-80.

[36] WEN B H, ZHU Y, SUBRAMANIAN R, et al. COVERAGE—a novel database for copy-move forgery detection[C]//Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE Press, 2016: 161-165.

[37] CARVALHO T J D, RIESS C, ANGELOPOULOU E, et al. Exposing digital image forgeries by illumination color classification[J]. IEEE Transactions on Information Forensics and Security, 2013, 8(7): 1182-1194.

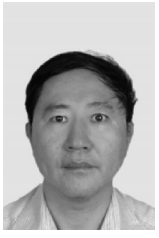
[38] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//European Conference on Computer Vision. Berlin: Springer, 2014: 740-755.

[作者简介]

谭舜泉（1980– ），男，广东湛江人，博士，深圳大学教授，主要研究方向为多媒体取证、隐写分析、深度学习等。

廖桂樱（1999– ），女，广西钦州人，深圳大学硕士生，主要研究方向为多媒体取证、深度学习。

彭荣煊（1998– ），男，广东揭阳人，深圳大学博士生，主要研究方向为多媒体取证、强化学习、深度学习。



黄继武（1962– ），男，广东揭阳人，博士，深圳北理莫斯科大学教授，主要研究方向为多媒体取证与安全、多媒体信号处理、信息隐藏等。