

用于图像篡改检测的混合 Transformer 网络

李树原¹⁾, 严彩萍^{1)*}, 李红²⁾

¹⁾ (杭州师范大学信息科学与技术学院 杭州 311121)

²⁾ (杭州启源视觉科技有限公司 杭州 311121)

(yancp@hznu.edu.cn)

摘要: 目前已有许多基于卷积神经网络的框架用于图像拼接伪造检测, 然而由于篡改区域的尺度是变化的, 现有的大多数方法都不能获得令人满意的性能, 尤其是对于大尺度的对象. 为了获得准确的篡改定位结果, 提出一种混合 Transformer 网络, 其将自注意力和交叉注意力结合到 U²-Net 中, 用于图像拼接伪造检测. 首先在编码器的最后一个模块应用自注意力捕获长距离语义信息依赖关系, 使网络能够更完整地定位大尺度篡改区域; 然后在跳跃连接中设计一个交叉注意力模块, 在高层语义信息的指导下增强低层特征图, 并过滤非语义特征, 实现了更精细的空间恢复. 混合网络中结合 Transformer 的自注意力和交叉注意力的优点, 能够从不同的尺度捕获更多的语义信息和空间依赖性, 即将卷积和 Transformer 融合在一起, 可以定位不同大小的拼接篡改区域, 而无需对大量图像进行预训练. 在 Casia2.0 和 Columbia 这 2 个公共数据集上, 与 4 种传统方法和 6 种深度学习方法进行实验的结果表明, 所提网络取得了更优的性能.

关键词: 图像拼接篡改检测; 篡改区域定位; 卷积神经网络; 自注意力; 交叉注意力

中图法分类号: TP391.41

DOI: 10.3724/SP.J.1089.2024.20099

A Hybrid Transformer Network for Image Splicing Forgery Detection

Li Shuyuan¹⁾, Yan Caiping^{1)*}, and Li Hong²⁾

¹⁾ (College of Information Science and Technology, Hangzhou Normal University, Hangzhou 311121)

²⁾ (Hangzhou Invision Technology Co., Ltd., Hangzhou 311121)

Abstract: In the past few years, many frameworks based on convolutional neural networks have been proposed for image splicing forgery detection. However, most of the existing algorithms can not obtain satisfactory performance due to tampered areas with various sizes, especially for objects with large-scale. In order to obtain an accurate forgery localization result, a hybrid Transformer architecture, which integrates both self-attention and cross-attention into U²-Net, is proposed for image splicing forgery detection. Specifically, self-attention is applied at the last block of encoder to capture long-range semantic information dependencies, so that the network can more completely locate large-scale tampered areas. Meanwhile, in the skip connections, a cross-attention module is designed to enhance the low-level feature maps with the guidance of high-level semantic information, filter out non-semantic features, and achieve more refined spatial recovery. Therefore, the hybrid model, which combines both advantages of self- and cross-attention from Transformer, has the ability to capture more context information and spatial dependencies from different scales. That is to say, the proposed method, fusing the convolution and Transformer together, can locate spliced forgeries with various sizes without requiring pre-training on a large number of images. Compared with four traditional

收稿日期: 2022-12-05; 修回日期: 2023-04-28. 基金项目: 国家自然科学基金(61902102); 浙江省自然科学基金(LQ19F020004).
李树原(1998—), 男, 硕士研究生, 主要研究方向为图像篡改检测; 严彩萍(1987—), 女, 博士, 副教授, 硕士生导师, CCF 会员, 论文通信作者, 主要研究方向为计算机视觉、图像篡改检测; 李红(1985—), 男, 博士, 主要研究方向为计算机视觉、图像处理.

methods and six new deep learning methods based on Casia2.0 and Columbia, the method in this paper achieves better performance.

Key words: image splicing forgery detection; tampered region localization; convolutional neural network; self-attention; cross-attention

随着现代移动设备的快速发展, 数字图像的生成和传递变得非常容易; 同时, 图像编辑软件操作简单, 普通人也能很轻易地对图像进行加工和修改. 通常, 人们修改图像是为了美化和娱乐, 然而一些篡改的图像可能被恶意滥用, 对社会和国家造成负面影响, 因此检测篡改图像变得越来越重要. 本文的研究重点是图像拼接篡改检测. 图像拼接是将图像的一部分复制并粘贴到另一个图像中, 以合成新图像. 在粘贴篡改区域的过程中可以执行一些后处理操作, 如模糊和平滑, 以隐藏篡改痕迹. 目前, 对图像篡改区域的检测分为基于传统特征提取的检测方法和基于卷积神经网络(convolutional neural networks, CNN)的检测方法 2 类.

大部分传统的检测方法通过提取特定的图像指纹进行检测, 如彩色滤波阵列插值^[1]、传感器噪声^[2]和光照不一致性^[3]. 当篡改涉及插入对比度高的物体时, 尤其是在明亮平滑的背景上, 这种传统的检测方法甚至可以检测出小物体并准确地划定其边界. 然而, 这种传统的检测方法在检测更复杂和微妙的篡改区域时会出现问题, 尤其是在纹理丰富的区域, 因为这些区域存在许多边缘会干扰分割; 此外, 该类方法只能检测特定的一种图像指纹, 当图像中的特定指纹不存在或不明显时检测就会失败, 且特定的图像指纹会受到图像模糊、JPEG 压缩和下采样等后处理效果的影响, 降低传统检测方法的检测结果.

基于局部性及其平移不变性, CNN 在计算机视觉领域取得了巨大的成功, 如图像分类^[4]、图像语义分割^[5]和图像目标检测^[6]. 因此, 人们提出了许多基于 CNN 的篡改检测方法, 它们表现出比传统方法更好的性能. 这些基于 CNN 的篡改检测方法可以同时提取多种图像指纹, 弥补了传统方法依赖单一图像属性、缺乏泛化能力和鲁棒性差的缺陷. Zhou 等^[7]基于 Faster Region-CNN, 使用 RGB 流和噪声流检测篡改图像的篡改区域, 使用富模型滤波器层从篡改图像中提取噪声特征, 可以使模型捕获篡改区域和真实区域之间的噪声不一致性; 该方法的局限性在于它只能实现区域级篡改定位结果. 为了获得像素级的结果, 一种可能的解

决方案是使用全卷积网络. Bunk 等^[8]提出一种混合 CNN-LSTM 模型, 可以通过学习共享边界中篡改区域和非篡改区域之间的空间结构进行篡改定位; Chen 等^[9]提出一种用于图像篡改检测的混合特征和语义增强网络(hybrid features semantic reinforcement network, HFSR-Net), 采用重采样特征的长-短期记忆, 可以从图像块捕获篡改轨迹. 然而, 上述方法都是通过先将图像分块来进行图像篡改检测, 意味着它们只关注局部区域而忽略了子区域之间的关系, 因此它们的定位性能取决于块的大小.

为了解决这个问题, 人们提出一些端到端网络, 将像素级篡改定位问题作为图像分割或目标检测任务. Bi 等^[10]提出一种环形残差 U 网络(ringed residual U-Net, RRU-Net), 通过残差传播和残差反馈模块捕获篡改区域和非篡改区域之间的特征, 增强了 CNN 的学习模式; 然而该网络不善于提取全局特征, 导致对大尺度篡改区域的定位效果较差, 误检率较高. Bi 等^[11]还提出一种多任务小波校正网络, 可以学习更全面和更具代表性的特征; 但该网络仍然没有考虑全局特征, 可能导致定位大尺度篡改区域时性能较差.

随着注意力机制的流行, 一些方法将注意力机制添加到 CNN 中, 以获得更好的性能. Rao 等^[12]提出一种深度 CNN, 其中结合了基于多语义条件随机场(conditional random fields, CRF)的注意模型, 用于图像篡改检测和定位, 该模型生成的注意力图有助于抑制噪声和突出信息区域, 并引导网络在篡改边界周围提取更具代表性的特征; Liu 等^[13]提出一种用于图像篡改检测和定位的渐进式空间通道相关网络(progressive spatio-channel correlation network, PSCC-Net), 采用高分辨率网络作为主干网络, 并设计了一个空间通道相关模块, 对提取的特征计算空间和通道注意力, 可以提高检测准确性, 并且能使网络提升鲁棒性.

然而, 已有的大多数工作忽略了篡改区域的大小是变化的, 并且难以定位不同大小的篡改区域. 由于卷积运算的固有局部性, 基于 CNN 的方法很难学习显式的全局语义信息关系, 并且难以联合利用局部和全局特征; 因此, 大多数基于

CNN 的检测方法只能处理有限的尺度变化. 此外, 这些方法在定位大尺度篡改区域时可能会出现定位不完整或高误检率等问题.

为了在不同尺度下准确地定位篡改区域, 本文提出一种 U 形混合 Transformer 网络, 其融合卷积和自注意力机制的优点, 用于图像拼接篡改检测和定位. 在编码过程中, 通过混合不同大小的感受野提取篡改图像的特征; 自注意力模块使用编码器末端语义特征之间的全局交互, 显式地建模完整的语义信息. 此外, 在跳跃连接中设计一个交叉注意力模块, 在解码器中实现更精细的空间信息恢复, 并提高预测结果的正确性. 经过自注意力模块后的特征输入解码器进行解码, 利用学习的特征估计最终的篡改掩模, 它可以同时获得局部和全局信息.

(1) 本文提出了一种 U 形混合 Transformer 网络, 它将自注意力和交叉注意力整合到 U^2 -Net 中, 能够从不问尺度捕获更多的语义信息和空间相关性.

(2) 本文在跳跃连接中设计了一个交叉注意力模块来滤除非语义特征, 在高级语义信息的指导下增强了通过跳跃连接的低级特征图, 并在解码器中实现精细的空间恢复, 从而最终提高预测结果的正确性.

(3) 在编码器的最后一个模块应用自注意力, 以结合卷积和自注意力机制的优点. 因此, 本文的

网络模型可以利用卷积的归纳偏差来避免大规模预训练, 以及 Transformer 学习建模显式长距离语义信息依赖的能力.

本文网络将卷积和 Transformer 融合在一起, 可以定位各种尺度的拼接篡改区域. 在 Casia2.0 和 Columbia 这 2 个图像篡改公共数据集上的实验结果表明, 本文网络取得了优于对比方法的性能.

1 本文网络

本文网络可以在像素级定位篡改区域, 其结构如图 1 所示. (1) 考虑到 Transformer 的强大能力, 将自注意力结合到编码器的最后一个模块来学习图像指纹, 可以更好地融合 Transformer 和卷积的优势, 即建模长距离特征依赖关系的能力和特征学习能力. (2) 为了解决最大池化导致图像语义信息丢失的问题, 设计一个交叉注意力模块来过滤编码和解码网络之间的非语义特征, 在解码器中实现更精细的空间恢复, 并提高预测结果的正确性. (3) 解码器学习融合从低分辨率到高分辨率的所有特征图, 并在像素级预测篡改区域.

1.1 编码器设计

当前, 大多数用于图像拼接篡改定位的神经网络使用传统的编码器-解码器提取特征, 然而由于大多数网络忽略了篡改区域的大小是变化的,

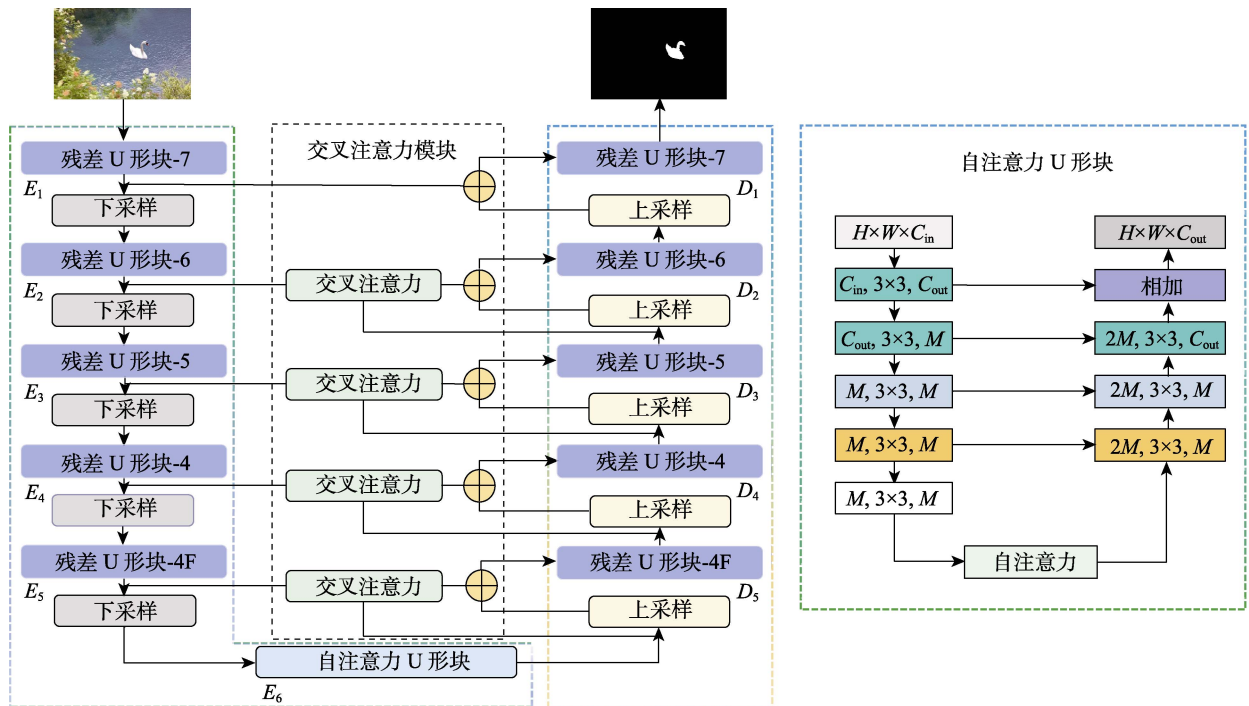


图 1 本文网络结构

因此只能处理有限的尺度变化. 与传统模型相比, U^2 -Net^[14]的一个显著改进是残差 U 形块能够混合不同大小的感受野, 即可以利用这一优点更好地处理任意大小的篡改区域问题. 因此, 本文使用 U^2 -Net 作为主干网络. 此外, 为了解决基于 CNN 的检测网络缺乏显式建模全局信息的能力的问题, 在编码器中应用自注意力模块捕获长距离语义信息.

本文设计一个 2 级嵌套 U 形混合 Transformer 网络, 对称结构可用于学习如何提取和编码多尺度语义信息. 在内层使用残差 U 形块, 能在不降低特征图分辨率的情况下提取级内多尺度特征. 此外, 嵌套的 U 形结构允许网络获得更高分辨率的特征图, 而不会显著增加内存和计算成本. 本文网络由 6 级编码器和 5 级解码器 2 个主要组件组成. 在编码器中, 分别使用残差 U 形块-7、残差 U 形块-6、残差 U 形块-5 和残差 U 形块-4, 其中, 7, 6, 5 和 4 表示残差 U 形块的高度 L , L 通常基于输入特征图的空间分辨率来配置. 对于具有更大高度和

宽度的特征图, 使用更大的 L 来捕获更多的大尺度信息. 如图 1 所示, 与 E_4 相比, 在 E_5 阶段, 池化和上采样操作由扩展卷积代替, 本文将该块命名为残差 U 形块-4F, 其中的所有中间特征图都具有与其输入特征图相同的分辨率; 在 E_6 阶段, 通过在原始残差 U 形块-4F 中添加自注意力模块, 形成自注意力 U 形块.

1.1.1 残差 U 形块

由于在普通卷积块中无法获得多尺度特征, 而使用扩展卷积虽然可以通过扩展感受野来提取局部和非局部多尺度特征, 但在具有原始分辨率的特征图上进行多次卷积需要太多计算和存储资源. 因此, 为了解决上述问题, 本文提出一种残差 U 形块捕获每个阶段的多尺度特征. 具有不同 L 值的 5 个残差 U 形块- L (C_{in}, M, C_{out}) 的结构如图 2 所示, 其中, L 表示编码器中的层数, C_{in} 和 C_{out} 表示输入和输出通道, M 表示残差 U 形块内部层中的通道数.

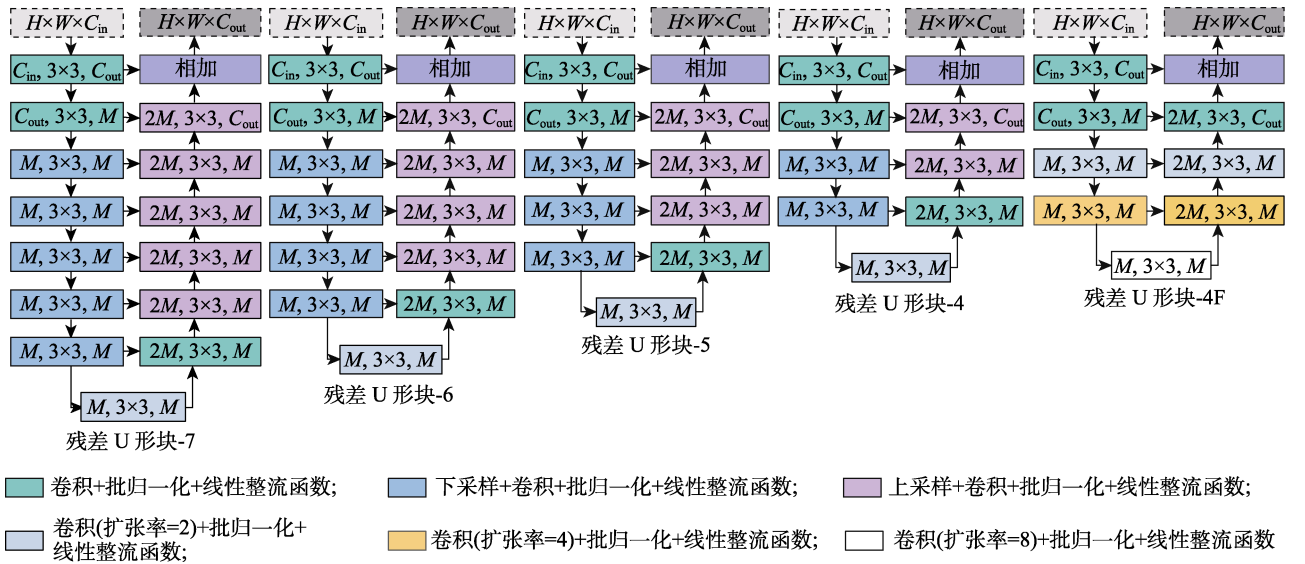


图 2 剩余 U 形块示意图

残差 U 形块由 3 个部分组成:

(1) 输入卷积层. 通过 3×3 的卷积核对输入特征图 $\mathbf{x} \in \mathbb{R}^{H \times W \times C_{in}}$ 进行变换, 以获得具有 C_{out} 通道的中间特征图 $F_1(\mathbf{x})$. 其中, H 和 W 分别表示特征图高度和宽度, 卷积层用于提取当前特征图的局部特征.

(2) 高度为 L 的类似 U-Net 对称编码器-解码器结构. 以中间特征图 $F_1(\mathbf{x})$ 为输入, 学习提取和编码多尺度语义信息 $U(F_1(\mathbf{x}))$. 该组件通过上采样、级联和卷积将多尺度特征编码为高分辨率特征

图, 减少特征丢失, 提高了篡改定位的准确性.

(3) 通过残差连接将其与之前的特征图相加, 从而融合局部与多尺度信息.

1.1.2 自注意力 U 形块

局部和全局语义信息对于图像拼接篡改检测和定位都非常重要. 自注意力 U 形块建立在自注意力模块^[15]之上, 可用于对图像的长距离特征依赖关系进行建模. 通过将自注意力整合到残差 U 形块的底部构建自注意力 U 形块, 如图 1 的右侧所示, 将最高级别特征图中的每个元素连接起来以

访问全局感受野. 自注意力模块的操作如图 3 所示, 首先在输入特征图 $\mathbf{X} \in \mathbb{R}^{d \times H \times W}$ 上添加位置编码, 其中, d 表示通道数, 位置编码特别适合于在自注意力中捕获篡改区域之间的绝对和相对位置; 然后将 $\mathbf{X}$ 展平并转置为大小为 $n \times d$ 的序列, 其中, $n = H \times W$; 再使用 3 个 1×1 卷积来投影 $\mathbf{X}$ 得到 3 个矩阵 $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{d \times n}$, 其中, $\mathbf{Q}$ 表示查询矩阵, $\mathbf{K}$ 表示键矩阵, $\mathbf{V}$ 表示值矩阵; 最后, 自注意力的输出是一个缩放点积, 公式为

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V} = \mathbf{A}\mathbf{V}.$$

其中, $\mathbf{A} \in \mathbb{R}^{n \times n}$ 被称为注意力矩阵或相似性矩阵, 注意力矩阵被用作权重, 以考虑 $\mathbf{Q}$ 和 $\mathbf{K}$ 之间的所有交互, 如图 3 所示. 这样, 自注意力本质上具有全局感受野, 擅长捕获长距离语义信息交互. 在拼接篡改定位任务中, $\mathbf{Q}, \mathbf{K}$ 和 $\mathbf{V}$ 的维度大小相同, 对应于图 3 中 $\mathbf{X}$ 所示的最高级别特征图的不同嵌入. 嵌入矩阵表示为 $\mathbf{W}_q, \mathbf{W}_k$ 和 $\mathbf{W}_v$.

1.2 交叉注意力模块

为了增强通过跳跃连接的低层特征图, 并在

解码器中实现精细的空间恢复, 本文在跳跃连接中设计了交叉注意力模块, 如图 4 所示. 注意, 本文没有在原始分辨率特征图上应用交叉注意力,

因为在网络非常浅的层中添加交叉注意力模块不仅不会提升性能, 反而会引入额外的计算. 为了保持跳跃连接的高分辨率信息, 利用交叉注意力模块在高级语义信息的指导下增强低层特征图. 交叉注意力模块的操作过程如图 4 所示. 首先在低层特征图 $\mathbf{U} \in \mathbb{R}^{d \times 2H \times 2W}$ 上添加位置编码; 然后将特征图 $\mathbf{U}$ 下采样一次, 并作为交叉注意力模块的值矩阵. 类似地, 在高级特征图 $\mathbf{N} \in \mathbb{R}^{2d \times H \times W}$ 添加位置编码, 并且用 1×1 卷积改变 d 之后作为查询矩阵和键矩阵; 然后, 使用矩阵乘法和 Softmax 函数计算交叉注意力矩阵 $\mathbf{A} \in \mathbb{R}^{n \times n}$, 其中, $n = H \times W$. 通过 ReLU 激活函数重新缩放计算的权重值, 得到的结果在图 4 中表示为 $\mathbf{S}$, 它可以看作一个滤波器; 其中, 低幅度元素表示要减少的噪声或无关区域. 将 $\mathbf{U}$ 和 $\mathbf{S}$ 进行 Hadamard 乘积, 得到 $\mathbf{U}$ 滤除噪声和非语义特征的版本; 将该滤波操作的结果与高级特征图 $\mathbf{N}$ 连接, 通过交叉注意力模块, 可以

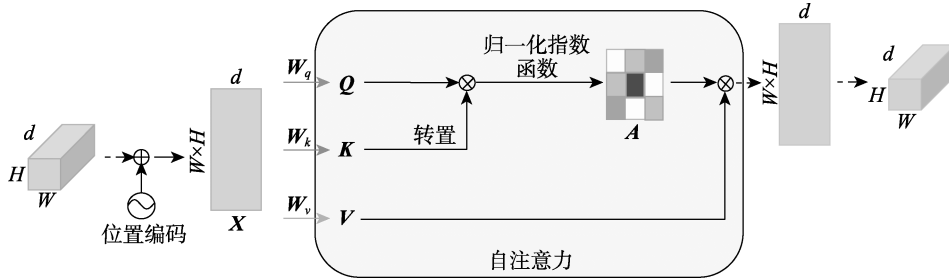


图 3 自注意力模块示意图

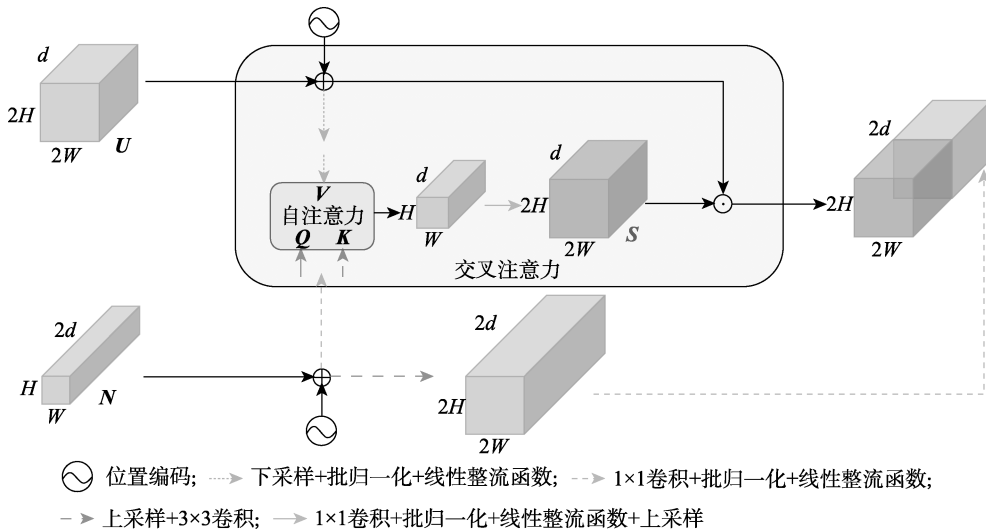


图 4 交叉注意力模块示意图

保留比普通跳跃连接更详细的信息, 提高检测性能.

1.3 解码器设计

解码器具有与其对称编码器类似的结构. 如在 D_5 中, 仍然使用扩展版本的残差 U 形块-4F, 其与 E_5 中使用的相似. 如图 1 所示, 每个解码器级将其前一级的上采样特征图和经过交叉注意力模块的输出特征图进行级联. 在解码器末端, 使用 3×3 卷积层和 Sigmoid 函数生成篡改区域的掩模.

本文网络可以捕获丰富的多尺度特征, 从而定位不同尺度的篡改区域.

2 实验结果与分析

2.1 实验设置

2.1.1 数据集

本文使用 Casia2.0^[16]和 Columbia^[17]这 2 个公共数据集, 分析和评估图像拼接篡改检测方法. 在 Casia 数据集中, 图像的篡改区域通过过滤和模糊进行精细篡改和后处理; 该数据集上的图像分辨率为 384×256 像素和 786×384 像素. Columbia 数据集中, 篡改区域大且简单. 本文从 Casia 数据集中选择 1828 组拼接篡改图像作为实验数据, 并将其随机分为 1622 个训练数据和 180 个测试数据. 同样地, 从 Columbia 未压缩数据集中选择 180 组拼接篡改图像, 并随机分为 160 个训练数据和 20 个测试数据.

2.1.2 实验参数

实验中, 首先将每幅图像的大小调整为 320×320 像素, 然后随机垂直翻转并裁剪为 288×288 像素. 本文网络采用 Adam 训练优化器进行训练, 学习率为 0.0001, 权重衰减设置为 0, batchsize 大小为 8, 模型迭代次数为 200. 验证集的比例为 0.1.

本文实验使用 I7 9700K CPU, 显卡为 GTX 3060 GPU, Python 版本为 3.8, Pytorch 版本为 1.8.1.

2.1.3 评估指标

为了在像素级评估图像拼接篡改检测方法的性能, 使用精确率、召回率和 F_1 值(F -measure)作为评估指标. 其中, 精确率表示检测到的区域是标定真值图像中真正被篡改的区域的概率; 召回率指标定真值图像中被篡改区域被正确检测到的概率; F_1 值将精确率和召回率结合为一个度量来综合评价检测方法的性能; 在实验中, 精确率、召回率和 F_1 值是测试集的平均值.

2.1.4 损失函数

通过二元交叉熵损失(binary cross entropy,

BCE)损失函数 L_{BCE} 对预测的结果进行优化, 计算公式为

$$L_{BCE} = - \sum_{(r,c)}^{(H,W)} [y_{(r,c)} \times \ln p_{(r,c)} + (1 - y_{(r,c)}) \times \ln(1 - p_{(r,c)})].$$

其中, (r, c) 表示像素坐标; $p_{(r,c)}$ 和 $y_{(r,c)}$ 分别表示输入图像预测的像素值和标签的像素值.

2.2 消融实验

为了验证注意力的影响, 使用 Casia 数据集作为训练集和测试集, 对 U^2 -Net 在逐步增加注意力下的性能进行评估, 结果如表 1 所示. 7 种网络如下:

(1) U^2 -Net^[14]. 未添加注意力的基础 U^2 -Net.

(2) U^2 -Net-base-SA-1. 用自注意力 U 形块替换 U^2 -Net 编码器最后一个残差 U 形块-4F.

(3) U^2 -Net-base-SA-2. 用自注意力 U 形块替代 U^2 -Net 的所有残差 U 形块-4F.

(4) U^2 -Net-base-SA-3. 在 U^2 -Net 编码器的最后一个残差 U 形块-4F 之后添加一个自注意力模块.

(5) U^2 -Net-base-SA-4. 在 U^2 -Net 编码器中的所有残差 U 形块编码器的最后 2 层添加一个自注意力模块.

(6) U^2 -Net-base-CA. 在 U^2 -Net 中添加 4 个交叉注意力模块.

(7) 本文网络. 自注意力和交叉注意力相结合的 U^2 -Net.

表 1 在 Casia 数据集上 7 种网络评估指标对比

网络	精确率	召回率	F_1 值
U^2 -Net ^[14]	0.603	0.663	0.593
U^2 -Net-base-SA-1	0.614	0.631	0.575
U^2 -Net-base-SA-2	0.613	0.575	0.553
U^2 -Net-base-SA-3	0.609	0.622	0.579
U^2 -Net-base-SA-4	0.725	0.734	0.694
U^2 -Net-base-CA	0.682	0.626	0.608
本文	0.809	0.746	0.735

注. 粗体表示最优值.

从表 1 可以看出, U^2 -Net-base-SA 和 U^2 -Net-base-CA 与原始 U^2 -Net 之间的性能逐渐提高, 清楚地表明了自注意力和交叉注意的有效性; 本文网络结合了自注意力和交叉注意力, 达到了最高的精确率、召回率和 F_1 值.

2.3 对比实验

采用不同检测方法进行拼接篡改检测性能的实验, 以验证本文网络的有效性, 包括 4 种传统的检测方法 ADQ^[18], CFA^[19], ELA^[20]和 NOI^[21], 以及

6 种基于深度学习的检测方法 RRU-Net^[10], U-Net^[22], ManTra-Net^[23], U²-Net^[14], MVSS-Net^[24] 和 Swin-Transformer^[25]. 在 Caisa 和 Columbia 数据集上, 通过精确率、召回率和 F_1 值评估 11 种方法的定位性能, 结果分别如表 2 和表 3 所示.

表 2 Caisa 数据集上 11 种网络篡改区域定位对比

方法	精确率	召回率	F_1 值
CFA ^[19]	0.144	0.697	0.202
ELA ^[20]	0.086	0.975	0.158
NOI ^[21]	0.149	0.992	0.258
ADQ ^[18]	0.402	0.585	0.476
U-Net ^[22]	0.734	0.629	0.626
U ² -Net ^[14]	0.603	0.663	0.593
RRU-Net ^[10]	0.678	0.586	0.586
ManTra-Net ^[23]	0.631	0.673	0.651
MVSS-Net ^[24]	0.758	0.736	0.719
Swin-Transformer ^[25]	0.433	0.425	0.429
本文	0.809	0.746	0.735

注: 粗体表示最优值.

表 3 Columbia 数据集上 11 种网络篡改区域定位对比

模型	精确率	召回率	F_1 值
CFA ^[19]	0.747	0.599	0.584
ELA ^[20]	0.316	0.961	0.475
NOI ^[21]	0.422	0.997	0.593
ADQ ^[18]	0.367	0.998	0.536
U-Net ^[22]	0.820	0.816	0.779
U ² -Net ^[14]	0.885	0.803	0.808
RRU-Net ^[10]	0.836	0.904	0.855
ManTra-Net ^[23]	0.716	0.549	0.621
MVSS-Net ^[24]	0.858	0.875	0.849
Swin-Transformer ^[25]	0.833	0.871	0.863
本文	0.915	0.878	0.872

注: 粗体表示最优值.

从表 2 和表 3 可以看出, 与基于深度学习的检测方法相比, 传统的检测方法的精确率和 F_1 值较差; CFA, NOI 和 ELA 得到了非常高的召回率, 因为它们几乎将整个图像检测为篡改区域; 本文网络在 2 个数据集上都达到了最高的精确率和 F_1 值; 本文网络、U-Net, RRU-Net 和 U²-Net 的参数数量分别为 1 397 421, 13 395 329, 4 097 249 和 1 131 181, 其中, U²-Net 使用的是较小的版本. 实验结果表明, 本文网络具有相对较少的参数数量, 取得了较好的检测性能. 进一步, 6 种方法的主观比较如图 5 所示, 对比方法中, 包括 2 种基于特征提取的传统检

测方法, 3 种基于深度学习的检测方法. 从测试数据集中选择 5 组数据, 其中, 第 1 组~第 4 组数据来自 Casia 数据集, 第 5 组数据来自 Columbia 数据集. 可以看出, ADQ 对 Casia 数据集的拼接篡改有一定的检测效果, 而对 Columbia 数据集的拼接篡改检测效果很差; ELA 对 Casia 数据集的拼接和篡改也有一定的检测效果, 与传统方法相比, 深度学习方法可以有效地定位 2 个数据集的篡改区域; U-Net 和 RRU-Net 仍然存在许多问题, 如错误分割、定位不成功和定位不完整; 本文网络融合了卷积和自注意力的优势, 因此可以更准确地定位不同大小的篡改区域.

2.4 攻击实验

在高斯噪声攻击和 JPEG 压缩 2 种攻击类型下, 评估本文网络的有效性和鲁棒性.

(1) 高斯噪声攻击. 在 Casia 数据集上, 对不同高斯噪声方差噪声攻击下图像的精确率、召回率和 F_1 值进行实验, 结果如图 6 所示. 可以看出, 随着高斯噪声方差的增大, 各评价指标都有不同程度的降低, 说明高斯噪声攻击对深度学习方法有一定影响, 但其性能仍优于传统的检测方法; 本文网络的精确率和 F_1 值远优于 6 种对比方法, 召回率略低于 RRU-Net, 且该网络的检测指数受到的影响最小. 实验结果表明, 本文网络对数据集的噪声攻击具有鲁棒性.

(2) JPEG 压缩攻击. 在 Casia 数据集上, 对不同 JPEG 压缩下图像的精确率、召回率和 F_1 值进行实验, 结果如图 7 所示. 可以看出, JPEG 压缩对图像篡改检测有一定的影响; 当质量因子从 100 降低到 50 时, CFA 和 ADQ 的精确率、召回率和 F_1 值会降低到非常低的值, 但对基于深度学习的方法的影响很小, 这是由于深度学习方法是一种基于图像内容特征的检测方法, 对压缩攻击具有很强的抵抗力, 因此影响较小; 本文网络仍然优于对比方法, 并且对数据集的压缩攻击具有鲁棒性.

3 结 语

本文提出一种混合 Transformer 网络. 该网络将自注意力和交叉注意力整合到 U²-Net, 不仅可以利用卷积的归纳偏差避免大规模预训练, 还拥有 Transformer 学习显式长距离语义信息依赖的能力. 与已有方法相比, 本文网络可以定位不同尺度的篡改区域. 为了增强通过跳跃连接传递的低层特征图, 并在解码器中实现精细的空间恢复, 设计

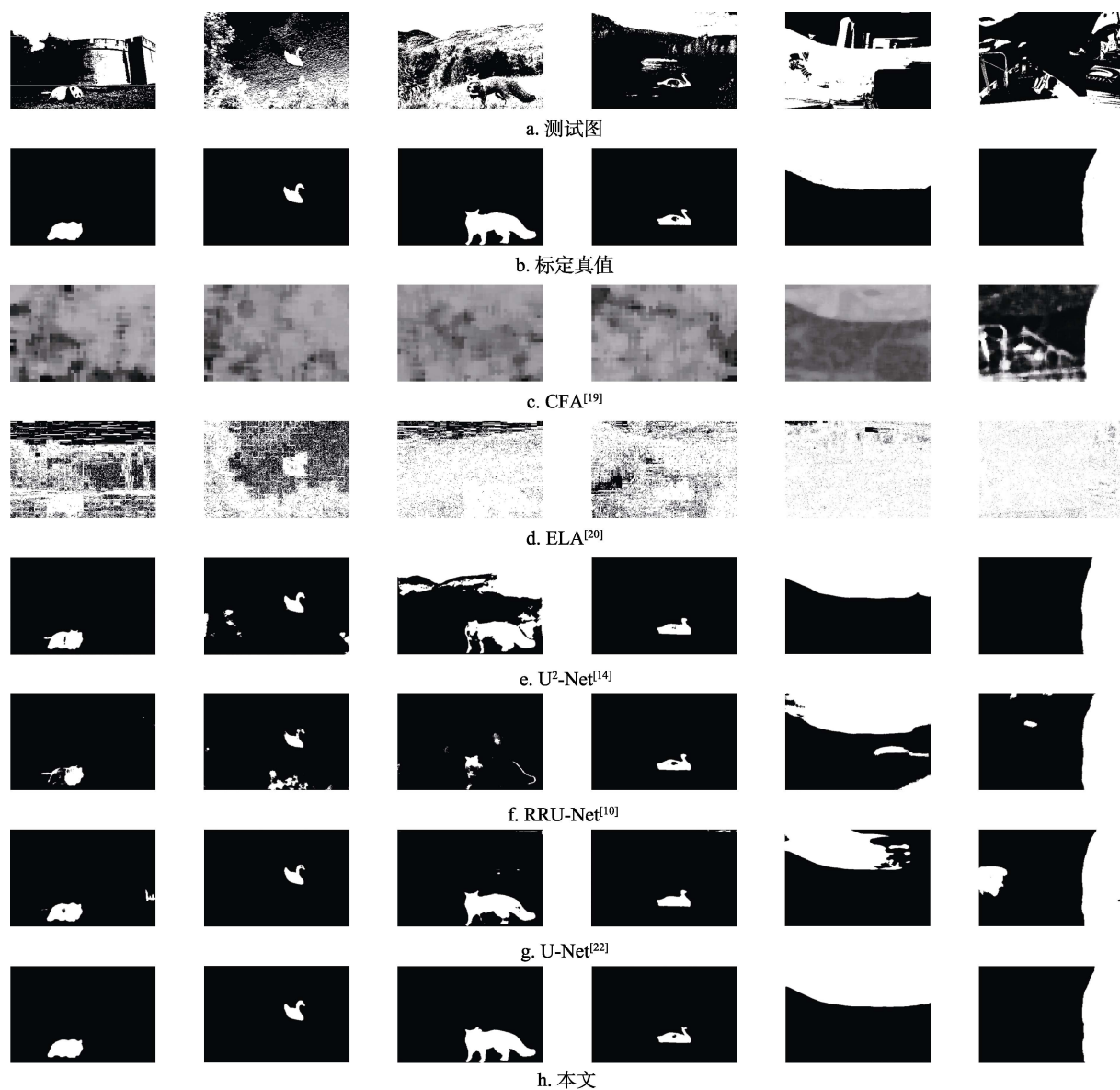


图 5 6 种方法的拼接篡改检测结果对比

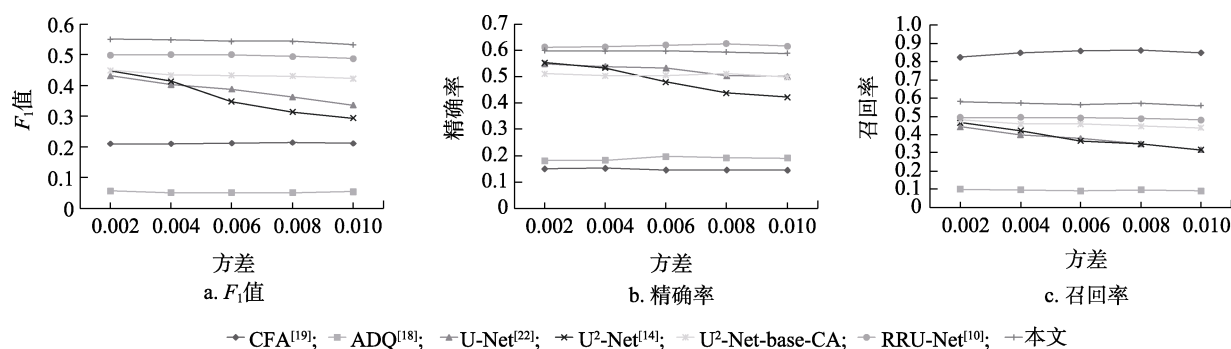


图 6 7 种方法在 Casia 数据集上添加高斯噪声的评价指标对比

了交叉注意力模块, 因此对于复杂图像的篡改更通用和有效. 在 2 个标准图像伪造数据集上的实验结果表明, 本文网络的性能优于对比方法. 但是, 该网络仍然存在一些待解决的问题: 虽然在多个

测试集上都有不错的表现, 但跨数据集测试的泛化性并不佳. 未来, 可以结合迁移学习的思想, 研究一种适用于跨数据集训练和检测的有效检测模型, 以提高模型的泛化能力.

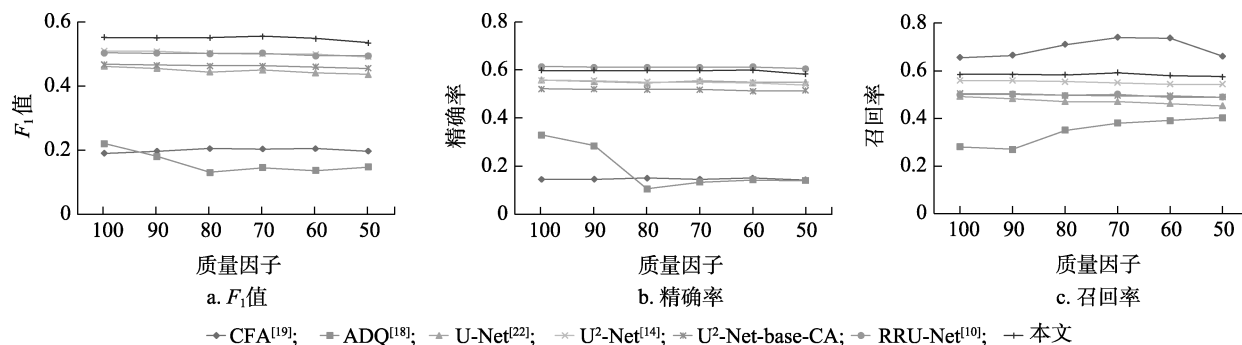


图 7 7 种方法在 Casia 数据集上 JPEG 压缩攻击的评价指标对比

参考文献(References):

- [1] Popescu A C, Farid H. Exposing digital forgeries by detecting traces of resampling[J]. *IEEE Transactions on Signal Processing*, 2005, 53(2): 758-767
- [2] Lukas J, Fridrich J, Goljan M. Digital camera identification from sensor pattern noise[J]. *IEEE Transactions on Information Forensics and Security*, 2006, 1(2): 205-214
- [3] 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
- [4] He K M, Zhang X Y, Ren S Q, *et al.* Deep residual learning for image recognition[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2016: 770-778
- [5] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2015: 3431-3440
- [6] Redmon J, Farhadi A. Yolov3: an incremental improvement[OL]. [2022-12-05]. <https://arxiv.org/abs/1804.02767>
- [7] Zhou P, Han X T, Morariu V I, *et al.* Learning rich features for image manipulation detection[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2018: 1053-1061
- [8] Bunk J, Bappy J H, Mohammed T M, *et al.* Detection and localization of image forgeries using resampling features and deep learning[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2017: 1881-1889
- [9] Chen H P, Chang C Q, Shi Z N, *et al.* Hybrid features and semantic reinforcement network for image forgery detection[J]. *Multimedia Systems*, 2022, 28(2): 363-374
- [10] Bi X L, 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*. Los Alamitos: IEEE Computer Society Press, 2019: 30-39
- [11] Bi X L, Zhang Z P, Liu Y B, *et al.* Multi-task wavelet corrected network for image splicing forgery detection and localization[C] // *Proceedings of the IEEE International Conference on Multimedia and Expo*. Los Alamitos: IEEE Computer Society Press, 2021: 1-6
- [12] Rao Y, Ni J Q, Xie H. Multi-semantic CRF-based attention model for image forgery detection and localization[J]. *Signal Processing*, 2021, 183: Article No.108051
- [13] 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
- [14] Qin X B, Zhang Z C, Huang C Y, *et al.* U²-net: going deeper with nested U-structure for salient object detection[J]. *Pattern Recognition*, 2020, 106: Article No.107404
- [15] Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need[C] // *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook: Curran Associates Inc., 2017: 6000-6010
- [16] Dong J, Wang W, Tan T N. CASIA image tampering detection evaluation database[C] // *Proceedings of the IEEE China Summit and International Conference on Signal and Information Processing*. Los Alamitos: IEEE Computer Society Press, 2013: 422-426
- [17] Hsu Y F, Chang S F. Detecting image splicing using geometry invariants and camera characteristics consistency[C] // *Proceedings of the IEEE International Conference on Multimedia and Expo*. Los Alamitos: IEEE Computer Society Press, 2006: 549-552
- [18] Lin Z C, He J F, Tang X O, *et al.* Fast, automatic and fine-grained tampered JPEG image detection via DCT coefficient analysis[J]. *Pattern Recognition*, 2009, 42(11): 2492-2501
- [19] 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
- [20] Krawetz N. A picture's worth... [OL]. [2022-12-05]. <https://www.hackerfactor.com/papers/bh-usa-07-krawetz-wp.pdf>
- [21] Mahdian B, Saic S. Using noise inconsistencies for blind image forensics[J]. *Image and Vision Computing*, 2009, 27(10): 1497-1503
- [22] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[C] // *Proceedings of the 18th International Conference on Medical Image Comput-*

- ing and Computer-Assisted Intervention. Heidelberg: Springer, 2015: 234-241
- [23] 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. Los Alamitos: IEEE Computer Society Press, 2019: 9535-9544
- [24] Dong C B, Chen X R, Hu R H, *et al.* MVSS-Net: multi-view multi-scale supervised networks for image manipulation detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3): 3539-3553
- [25] Liu Z, Lin Y T, Cao Y, *et al.* Swin transformer: hierarchical vision transformer using shifted windows[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 9992-10002