智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于深度学习的图像篡改检测方法综述

张汝波, 蔺庆龙, 张天一

引用本文:
张汝波, 蔺庆龙, 张天一. 基于深度学习的图像篡改检测方法综述[J]. 智能系统学报, 2025, 20(2): 283-304.
ZHANG Rubo, LIN Qinglong, ZHANG Tianyi. A review of image tampering detection methods based on deep learning[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(2): 283-304.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202403004

DOI: 10.11992/tis.202403004
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20250123.1117.002

基于深度学习的图像篡改检测方法综述

张汝波[1], 蔺庆龙[1], 张天一[2]
（1. 大连民族大学 机电工程学院, 辽宁 大连 116600; 2. 北京航空航天大学 网络空间安全学院, 北京 100191）

摘　要: 随着数字图像编辑工具的普及, 图像篡改变得越来越容易, 大量被篡改后的虚假图像通过网络和社交媒体进行传播, 这对法律、新闻媒体和科学研究等领域的真实性和可信度构成了威胁。图像篡改检测的目的是检测和定位篡改图像中的篡改区域, 以保护图像的可信度。本文对基于深度学习的篡改检测方法进行了回顾总结。首先, 介绍了目前图像篡改检测领域的研究现状。其次, 对近 5 年的深度学习方法进行了分类整理。然后, 介绍了主要的数据集和评价指标, 以及各种方法的性能对比。最后, 探讨了目前篡改检测方法的局限性并对未来的发展方向进行了展望。

关键词: 深度学习; 图像篡改检测; 计算机视觉; 卷积神经网络; 图像处理; 图像取证; 图像伪造; 伪造检测
中图分类号: TP39　　**文献标志码:** A　　**文章编号:** 1673−4785(2025)02−0283−22

中文引用格式: 张汝波, 蔺庆龙, 张天一. 基于深度学习的图像篡改检测方法综述 [J]. 智能系统学报, 2025, 20(2): 283–304.
英文引用格式: ZHANG Rubo, LIN Qinglong, ZHANG Tianyi. A review of image tampering detection methods based on deep learning[J]. CAAI transactions on intelligent systems, 2025, 20(2): 283–304.

A review of image tampering detection methods based on deep learning

ZHANG Rubo[1], LIN Qinglong[1], ZHANG Tianyi[2]
(1. College of Mechanical & Electronic Engineering, Dalian Minzu University, Dalian 116600, China; 2. School of Cyber Science and Technology, Beihang University, Beijing 100191, China)

Abstract: With the increasing popularity of digital image editing tools, image tampering has become much easier. A large number of tampered false images are now circulating on the Internet and social media, threatening the authenticity and credibility of critical domains such as law, journalism, and scientific research. Image tampering detection aims to identify and locate altered areas within tampered images, thereby safeguarding their credibility. This paper provides a comprehensive review of deep learning-based methods for image tampering detection. First, it introduces the current research status in this field. Next, it classifies deep learning approaches developed over the past five years. The paper also highlights the main datasets and evaluation metrics used, along with a performance comparison of various methods. Finally, it discusses the limitations of current tampering detection methods and offers insights into future development directions.

Keywords: deep learning; image tempering detection; computer vision; convolutional neural network; image processing; image forensic; image forgery; forgery detection

随着先进数字图像处理技术的快速普及和发展[1-3] 以及智能手机和各种图像编辑软件的广泛使用, 图像很容易被编辑篡改, 从而产生具有误导性或欺骗性的视觉内容。这种图像的易篡改性不仅引起了公众对媒体真实性的担忧, 而且在法律和科学研究领域也带来了一系列的信任和安全问题[4]。因此, 有必要发展有效的图像篡改检测技术, 以辨别原始图像内容和被篡改图像内容之间的差异, 确保图像的真实性。在早期传统的图像篡改检测方法中[5-6], 主要是依据成像设备的固

收稿日期: 2024−03−04.　**网络出版日期:** 2025−01−23.
基金项目: 国家自然科学基金项目 (62202024); 中央高校基本科研业务费专项资金项目 (501QYJC2024139006).
通信作者: 张天一. E-mail: zhang_tianyi@buaa.edu.cn.

有属性[7]、图像的内在统计特征[8]以及图像的篡改痕迹[9]这 3 种方法来进行分析检测的，但这类方法容易受到噪声和图像压缩等因素的干扰，从而导致检测定位的结果不够精准。另一方面，传统的图像篡改检测方法需要手工设计进行特征提取，这需要依赖领域专家的经验和知识，并可能会受到主观因素的影响，导致提取的特征不够客观和全面[10-11]。随着深度学习在计算机视觉领域的广泛应用，越来越多的研究者开始将深度学习技术引入到图像篡改检测领域中来。相比传统的图像篡改检测方法，基于深度学习的图像篡改检测方法可以通过训练自动学习图像的特征，无需手工提取特征，大大减少了人为因素的干扰，并且在检测定位的准确率上也取得了更优秀的表现[12-14]。此外，深度学习方法具有较高的鲁棒性和扩展性，能够适应不同的图像篡改检测任务[15-16]。

基于目前图像篡改检测领域的研究现状，本文考虑对基于深度学习的图像篡改检测定位方法进行整理。目前已有许多关于图像篡改检测的综述工作[4,14,17-19]，与现有工作相比，本文的创新点如下：首先，本文重点整理了图像篡改检测领域近 5 年来发表在重要会议与期刊上的最新研究成果。由于近年来在篡改检测领域，深度学习方法相对传统方法已经具有显著优势，因此本文着重讨论基于深度学习的图像篡改检测方法；其次，本文基于全新的分类方法对图像篡改检测方法进行归纳梳理，重点提炼分析了方法中的共性关键技术，例如多流信息融合、多尺度特征融合、基于边缘信息等，这些共性关键技术对于不同的深度学习网络架构具有较高的普适性，对篡改检测精度的进一步提升有一定参考意义；最后，本文增加了新的分析视角，除通常在标准数据集下的场景外，本文增加了面向现实复杂场景下 (例如有损后处理、网络传输) 的研究工作总结，对于图像篡改检测方法在现实场景下的广泛应用具有一定的促进作用。

本文的后续内容安排如下：第 1 节介绍目前基于深度学习的篡改检测方法，并根据不同方法的特点进行分类总结。第 2 节介绍常用的数据集和评价指标，并对不同方法的定位性能进行对比。第 3 节指出目前各类篡改检测定位方法存在的问题并对未来的发展方向进行展望。第 4 节对本文的内容进行总结。

1　基于深度学习的图像篡改检测方法

图像篡改是指对数字图像进行恶意修改或伪造，以改变图像原本的语义内容从而达到欺骗、隐瞒或误导人们的目的。根据对图像进行篡改时的操作类型不同，篡改方式主要分为 3 种：拼接 (splicing)、复制移动 (copy-move) 和删除 (remove)。3 种篡改方式的例子如图 1 所示。

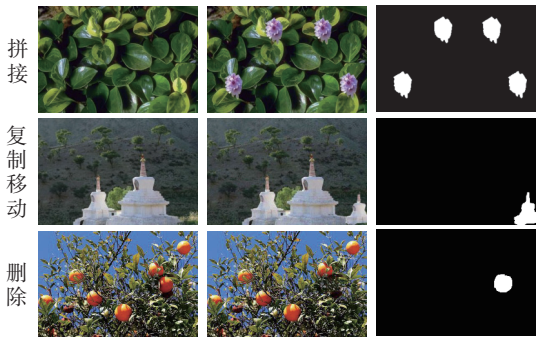

图 1　图像篡改方式

Fig. 1　Image tampering methods

近几年来，研究者提出了许多基于深度学习的篡改检测模型，这些方法利用深度神经网络的强大特征学习和表示能力，能够有效地检测图像中的篡改区域。本文按照图 2 的架构对基于深度学习的图像篡改检测方法进行归纳。首先按照基于多流信息融合、多尺度特征提取、边缘信息、对比学习等共性关键技术进行总结。该关键技术多数基于卷积神经网络展开，但对不同深度学习架构具有一定普适性。其次，本文对基于其他深度学习网络架构的方法进行总结，重点介绍基于 Transformer 架构的最新研究方法。最后，本文对面向现实复杂场景下 (例如有损后处理、网络传输) 的研究工作进行归纳。

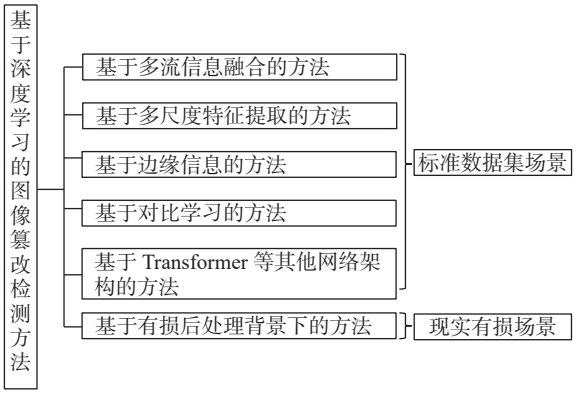

图 2　篡改检测方法分类

Fig. 2　Classification of tampering detection methods

1.1　基于多流信息融合的方法

多流信息融合是目前图像篡改检测领域常用的模型架构，如图 3 所示。多流信息融合方法是指在原有 RGB 域输入的基础上，增加了噪声域、频域等多种输入形式，每一项输入形式都构成信

息流从而达到多流信息融合共同优化篡改检测定位精度的目的。

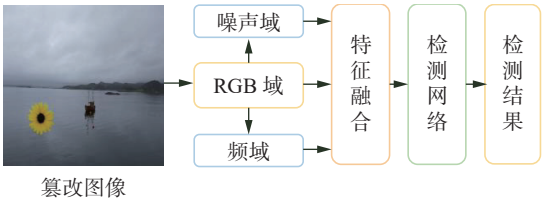

图 3　基于多流信息融合的方法

Fig. 3　Methods based on multi-stream information fusion

由于图像篡改检测不同于传统的语义对象检测，它更关注的是图像的篡改痕迹而不是图像的内容，因此如何从图像中提取出更多的篡改痕迹信息变得至关重要，如果仅使用 RGB 图片作为输入，可能无法提取出足够的篡改痕迹信息供模型学习。因此，研究者考虑使用多流信息融合的模型架构，通过在噪声域和频域中提取非语义信息，补充 RGB 图像中可能忽略的篡改痕迹，从而实现更精准的定位效果。

当前，在多流信息融合方法中，噪声域信息[20-21]是最常用的输入流之一。与 RGB 信息相比，噪声信息对篡改区域的变化更敏感，可以提供更多的篡改痕迹，从而能够更好地区分篡改区域和真实区域。Zhou 等[22]首先提出了采用 RGB 和噪声域进行双流特征提取的思想，通过 SRM(steganalysis rich model)[23]滤波器将 RGB 图像转换为噪声图。SRM 滤波器通过计算像素值与其邻近像素值之间的残差来建模噪声。初始 SRM 使用 30 个基本滤波器来收集局部噪声特征，但在实际实验过程中发现，仅仅使用 3 个内核就可以实现不错的性能，因此目前常用的做法都是仅使用 3 个滤波器，其权重如图 4 所示。

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

图 4　提取噪声的 3 个 SRM 滤波器内核

Fig. 4　3 SRM filter kernel for noise extraction

在主网络的设计上，使用 Faster R-CNN(fast region-based convolutional network method)[24]作为主网络架构，将 RGB 图像作为第 1 输入流，提取的噪声图作为第 2 输入流分别传入 2 个并行的特征提取网络中。使用 RPN(region proposal network) 网络层从 RGB 流提取生成候选区域，并根据提取的候选区域在双流分支中提取特征，最后通过双线性池融合来自 2 个流的特征，用于最终分类网络进行区分篡改类型。遗憾的是，由于该方法采用 Faster R-CNN 网络，因此无法实现精准的像素级别分类，只能给出大致篡改区域的边界框。Rao 等[25]同样采用提取噪声域信息的双流架构思想进行图像篡改检测定位。首先分别提取空间域和噪声残差域信息，随后将双流的特征进行连接融合，最后使用 8 个串联的残差模块提取融合特征中不同尺度的层次特征。在特征融合阶段，此方法先将双流的初级特征进行融合，而后再进行进一步的特征提取，文献 [22] 则是先提取深层特征，而后再进行融合。Chen 等[26]和 Li 等[27]以及 Liu 等[28]也都采用了噪声分支，并使用 Bayar Conv[29]来提取噪声信息。Bayar Conv 能够更好地抑制图像中的内容信息，使模型专注于捕获图像的篡改痕迹，从而更好地区分篡改区域和真实区域。对于 2 个分支的融合，Liu 等[28]设计了一种注意感知的特征融合模块，该模块通过构建一个位置注意力块来研究 RGB 特征和噪声特征的相关性并将它们融合为统一的特征图，而 Chen 等[26]和 Li 等[27]则采用了可训练的双重注意力 (dual attention, DA) 模块[30]，该模块在位置注意力模块的基础上增加了通道注意力模块。双重注意力模块能够更敏感地捕捉篡改痕迹，并根据具体任务学习到最优的特征，从而实现更优的特征融合。

上述提到的方法都仅使用一种噪声提取方法进行噪声特征提取，而有研究者发现联合使用各种噪声提取方法会取得更好的定位性能。Wu 等[31]提出了一种用于检测和定位具有异常特征的图像篡改定位方法。该方法包括 2 个子网络：特征提取网络和局部异常检测网络。在特征提取网络设计上，此方法将经典的 RGB Conv、SRM 滤波器和 Bayar Conv 进行组合使用。首先将 3 种不同的基础特征进行简单的融合，而后使用串联的卷积块进行深层次的特征提取。局部异常检测网络则由 3 个小阶段构成：1) 适应阶段，用于将篡改痕迹特征适配到异常检测任务中；2) 异常特征提取阶段，该阶段受到人类思维的启发，用于提取异常特征；3) 决策阶段，从整体上考虑异常特征并分类判断像素是否被伪造。大量实验证明，该方法具有相对良好的泛化能力与鲁棒性。受文献 [31] 的启发，Lin 等[32]同样设计了一个包含 SRM 滤波器、Bayar 滤波器和普通 RGB Conv 的组合块从 RGB 图像中提取噪声特征，但是在噪声信息的深层次特征提取上与上述方法存在着差异。上述方法都使用 CNN(convolutional neural network) 提取局部噪声特征，而没有探索全局噪

声特征，而文献 [32] 使用 Swin Transformer 来探索全局噪声特征和局部噪声特征，从而可以更全面地提取篡改痕迹。Das 等[33] 提出了一个多模态的特征提取模块用以提取不同领域的特征。除文献 [31-32] 中提到的 3 种方法外，还增加了改进的错误级别分析 (error level analysis, ELA) 模块[34] 以提取图像信号噪声和压缩伪影。ELA 曾用于定位 JPEG 图像中的压缩伪影，它通过比较图像与其压缩副本之间的像素差异来工作。如果图像包含来自不同来源的像素，这些像素会产生不同程度的压缩噪声。在该工作中，作者对图像进行 90% 压缩，然后计算原始图像与压缩图像之间的差异，以生成 ELA 输出。该 ELA 输出经过卷积层处理，再应用激活函数，得到 ELA 特征图并输入到编码器网络中。ELA 特征图能够反映篡改区域与真实区域之间像素级别差异，有助于最终的篡改检测定位。Guillaro 等[35] 提出了一种利用全方位线索进行图像篡改检测和定位的方法 TruFor。此方法提出了一种新颖的噪声特征提取技术，即 Noisrprint++，它是在 Noiseprint[36] 的基础上进行改进的。Noiseprint 可以通过捕获图像的噪声特征信息区分图像的拍摄设备是否相同，而 Noisrprint++不仅可以区分图像的拍摄设备是否相同，还可以捕获图像的历史编辑信息，从而得知图像经历过哪些编辑操作。除了上述贡献，TruFor 还提出了新颖的篡改检测框架，它包含 3 个输出：1) 全局的完整性分数用于评估图像的整体真实性，分数的高低对应着图像的真伪程度；2) 基于异常的定位图，用于可视化定位图像中可能被篡改的区域；3) 与异常图相辅相成的置信度图，提供了关于异常定位图中每个区域的置信度信息，用于辅助用户更准确地判断图像中的异常是否可信，从而提高伪造检测的可靠性。总的来说，这样的设计提供了一个更全面的视角来评估图像的真实性，不仅能够快速识别可能的伪造，还能精确地定位篡改区域。

除了利用噪声域信息，部分研究者还利用了频域信息来提供 RGB 域中不可见的篡改痕迹。同样的，在 RGB 域中不明显的视觉伪影，在频域中却常常很明显，于是通过融合 RGB 和频域信息能够获得更加丰富的篡改痕迹。Wang 等[37] 提出了一种结合频域信息和 RGB 信息的图像篡改检测定位方法。具体地，首先使用离散余弦变换 (discrete cosine transform, DCT) 将 RGB 图像转换到频域，利用高通滤波器提取高频分量，得到高频特征，然后将获得的高频特征与 RGB 特征拼接在一起作为模型的输入。Guo 等[38] 设计了一个基于细粒度分类的篡改检测定位框架，该框架通过 Color 块与 Frequency 块分别提取输入图像的 RGB 和频域特征，并且在 Frequency 块中应用高斯拉普拉斯算子 (Laplacian of Gaussian, LOG) 来抑制图像中的语义信息，进而放大篡改痕迹。Kwon 等[39] 设计了一个完整端到端的图像篡改检测网络，它包括 RGB 流、DCT 流和融合阶段。该方法使用并行的 HRNet(high resolution network) 同时处理 RGB 和 DCT 流。首先，RGB 图像被输入到压缩伪影学习模块中，提取 DCT 特征，这些特征能捕捉图像中的压缩伪影信息。与此同时，另一个流处理 RGB 图像的原始信息。2 个流的特征经过处理后，被分别上采样到相同的尺寸，然后融合在一起。这种融合过程结合了 RGB 图像的细节信息和 DCT 特征中的压缩伪影信息，从而提高了最终预测的准确性。与文献 [39] 类似，Zhang 等[40] 也同时结合了 RGB 和频域流，但在特征提取阶段存在不同。该方法还使用 ASPP(atrous spatial pyramid pooling) 模块来捕获不同尺度的信息，并使用通道和空间注意力的相互作用以增强特征。在特征融合阶段，也提出了新颖的软选择方法对 RGB 流和频域流的 2 个热图进行加权聚合，生成最终的加权热图 (预测图)。Gu 等[41] 提出了一种基于 DCT 的多任务图像篡改检测模型。该模型首先将 RGB 图像利用 DCT 转换到频域，而后利用滤波器分别提取高频 DCT 滤波图像和低频 DCT 滤波图像，而后通过参数共享的方式将低频、高频和原始图像输入到并行的共享编码器网络中，以学习不同频域类型的图像特征表示，其中高频特征有助于突出图像中的失真，而低频特征则对保留图像的整体内容起到重要作用。在编码器网络后，提出了扩张频率自注意力模块，它通过自注意机制对编码后的特征进行增强，以提取更具代表性的特征。

1.2　基于多尺度特征提取的方法

在图像篡改检测任务中，多尺度特征提取指的是从不同的尺度 (分辨率或大小) 上提取的图像特征，通过跳跃连接等形式将不同尺度特征进行整合，以获得更全面的特征表示。多尺度信息的架构示意如图 5 所示。通过使用多尺度特征提取方法，可以在不同层次上获取图像的细节信息和全局上下文信息，这有助于发现不同尺度下的篡改痕迹，并能够增强算法对不同尺度下篡改方法的鲁棒性。

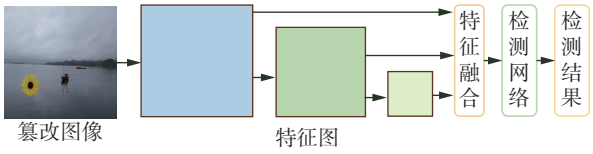

图 5　基于多尺度特征提取的方法

Fig. 5　Methods based on multi-scale feature extraction

Liu 等[42] 提出了一种基于 Atrous 卷积的对抗学习框架，通过权重共享的并行网络进行特征提取。为了适应篡改图不同尺度的篡改区域，设计了一种 ASPP 模块，用以捕捉不同尺度的信息。ASPP 包含 4 个并行的空洞卷积层，每个空洞卷积层具有不同的采样率。具体地，每一层的空洞卷积层后都有一个单独的卷积层、批量归一化层和 ReLU 层，这样的设计可以关注不同尺度的篡改区域，从而在不同的尺度上捕获细微的篡改痕迹。最后通过将这些并行的空洞卷积层分支融合在一起，综合考虑不同尺度的特征信息，生成最终的篡改检测定位掩码。Niloy 等[43] 同样采用了 ASPP 模块进行图像篡改检测定位。与文献 [42] 的多尺度方法有些许不同，此方法使用 4 个具有不同采样率的空洞卷积层对融合后的特征进行分层次的特征提取，并使用 1 个单独的 1×1 卷积层保留原始的尺度信息，从而获取更丰富的全局上下文信息，提高检测的准确性，而后将 ASPP 模块输出的多尺度特征传入到分割头中，生成最终的预测掩码。Zhong 等[44] 提出了一个端到端的 InceptionNet 网络来提取多尺度的特征。具体而言，通过 3 个串联的金字塔特征提取器 (pyramid feature extractor, PFE) 模块逐步地压缩特征图的深度和尺寸，生成 3 个由大到小的特征图，然后，通过 3 个与 PFE 相对应的特征融合模块 (feature correlation matching, FCM) 分别对不同尺度的特征进行统一融合，以便挖掘出不同尺度的特征信息。Guo 等[38] 提出了一个多层次细粒度的图像篡改检测定位方法。此方法提出了一个多分支的特征提取器，一共包含 4 个分支，每个分支生成特定分辨率的特征图，不同分辨率的特征图用以在不同尺度上进行细粒度的伪造类型分类，例如，如果仅需要简单地识别图片是否经过篡改，则仅需要最粗级别的分支，而如果需要更精细的篡改区域定位，并识别各个具体的伪造方法，则需要最精细的分支提供高分辨率的特征图。为了实现精准的像素级别定位，作者将定位模块添加到最精细的分支后，并且为了模拟像素在空间区域上的依赖性和相互作用，采用了包含自注意力机制的定位模块，其本质上模仿 Transformer 中的自注意力机制。此外，此方法不仅可以检测传统的篡改方式生成的篡改图，也可以对使用 GAN 或者 Diffusion Model 等方法的 DeepFake 篡改起到作用，这也是目前所提出篡改检测方法中的一个比较新的尝试。Hao 等[45] 利用全卷积网络 (fully convolutional networks, FCN) 作为特征提取的主干网络，分别生成 4 个不同尺度的特征图，然后对 4 个不同尺度的特征图分别应用 Transformer 网络的自注意力机制，对全局上下文以及不同尺度的补丁块的关系进行学习，最后将 4 个不同尺寸的特征图上采样到相同的尺寸后进行融合，通过步长为 1 且填充为 1 的 3×3 卷积获得预测掩码。Lin 等[32] 则设计了一个 ResNet(residual networks) 和 Transformer 相结合的框架，用 4 个串联的 ResNet 块来提出不同尺度上的 RGB 特征。对于噪声域信息，则采用 Swin Transformer 块进行全局和局部噪声特征的提取，具体地，Swin Transformer 块[46] 通过使用窗口自注意力机制 (window-based multi-head self-attention, W-MSA) 提取全局噪声特征，通过使用移动窗口自注意力机制 (shifted window-based MSA, SW-MSA)[47] 提取局部噪声特征，然后将不同 ResNet 块和 Swin Transformer 块提取的不同尺度上的特征进行融合，以获取更全面的特征表示。Das 等[33] 则基于 Unet 网络[48] 设计了一种改进的 Unet++网络框架，Unet++在编码器和解码器之间添加了多个跳跃连接，使得底层特征可以直接传递到解码器的对应层，从而提供更丰富的多尺度信息。此外，为了更好地利用不同尺度的信息和全局上下文信息，还提出了一种门控上下文注意力机制模块，该模块被添加到不同的 Unet++层中，根据学习特征的相对重要性来调节学习特征，具体包括 2 个阶段：全局上下文池化和注意力门控。首先通过全局上下文池化对不同层的特征进行全局信息聚合来识别长距离的依赖关系，接着利用注意力门控机制生成注意力系数矩阵来筛选和保留重要的特征，从而抑制不相关的特征信息。Liu 等[49] 则提出了一种渐进式的空间通道相关网络，主要包括 2 个部分：编码器部分是自上而下的路径 (top-down path)，使用 HRNet[50] 来提取输入图片的局部和全局特征，共包含 4 层分支，每一层提取不同尺度的特征；解码器部分是自下而上的路径 (bottom-up path)，用以从不同的尺度生成篡改图像的预测掩码图和预测分数。在每一层中都额外添加了常用的空间通道注意力模块对学习到的特征进行优化和聚合。值得注意的是，4 层路径中的每一层都会自上而下由粗到细

生成不同尺度的预测掩码，除最顶层外，其余的每一层预测掩码都作为下一层预测掩码的先验知识，这样的设计能够生成更为精细的预测掩码，提高定位的准确性。

1.3　基于边缘信息的方法

由于图像篡改区域和真实区域具有较高的相似性，导致篡改定位方法无法产生准确清晰的篡改边缘，通常情况下位于 2 个区域交界处的边缘区域会包含更多的篡改痕迹。因此，增强特征中边缘信息的表达，以及在监督信号中增加边缘监督任务，可以一定程度上提升篡改检测定位的精度。

目前的许多研究者也提出了相关的边缘信息篡改检测方法，其主要在 2 个方面进行研究：一部分研究者通过设计各类边缘增强模块以提取出更多的边缘特征，从而增强篡改定位的精度，如图 6 中方框①所示；另一部分研究者则通过真实值区域掩码生成边缘掩码，将边缘掩码也作为监督信号，设计边缘损失策略来监督指导模型更好地学习边缘特征，如图 6 中方框②所示。

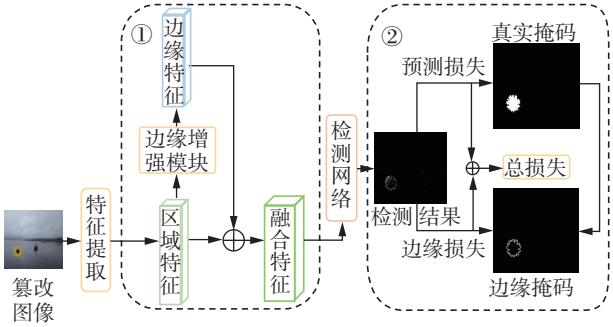

图 6　基于边缘信息的方法

Fig. 6　Methods based on edge information

Lin 等[32] 提出一个包含边缘解码分支与区域解码分支的图像篡改检测定位框架。通过边缘解码分支来增强图像中的篡改痕迹，以发现图像中隐藏的微妙边缘伪影。边缘解码分支由边缘增强模块、边缘上采样模块和边缘监督策略组成。边缘增强模块能够消除图像中篡改区域周围的边缘伪影，然后通过计算输入特征图与消除伪影后的特征图之间的残差，从而间接增强融合特征图中的边缘特征。边缘上采样模块则对边缘增强模块输出的边缘特征图进行上采样，提高特征图的分辨率，增强边缘特征，以便更好地检测和恢复图像中边缘细节。边缘监督策略是指通过边缘监督来强制边缘解码分支专注于检测细微的边缘伪影。作者使用 Dice 损失函数作为边缘监督的损失函数，这是因为篡改区域周围的边缘像素占比相对较少，Dice 损失能够忽略不重要的背景区域像素，从而更关注边缘像素，非常适用于边缘监督任务，其公式为

$$\mathrm{loss_e}(x)=1-\frac{2\cdot\sum_{i=1}^{H\times W}E(x_i)\cdot y_i}{\sum_{i=1}^{H\times W}E(x_i)^2+\sum_{i=1}^{H\times W}y_i^2}$$

式中：$E(x_i)$是边缘解码分支的预测结果，表示图像x中第i个像素被操纵边界的概率；y_i表示二进制边缘真实值 (0 代表非边缘，1 代表边缘)，指示第i个像素是否属于操纵区域的边界。除边缘监督外，还针对区域解码分支提出了相应的监督策略，使用二元交叉熵作为该分支的损失函数，其公式为

$$\mathrm{loss_r}(x)=-\frac{\sum_{i=1}^{H\times W}[y_i\ln R(x_i)+(1-y_i)\ln(1-R(x_i))]}{H\times W}$$

式中：$R(x_i)$表示区域分支的预测结果，y_i表示对应的掩码真实值像素，H、W分别代表输入图片的高和宽。最终的总损失为

$$\mathrm{loss_t}(x)=\gamma_e\cdot\mathrm{loss_e}(x)+\gamma_r\cdot\mathrm{loss_r}(x)$$

式中γ_e和γ_r分别是边缘和区域损失的权重，用以平衡不同损失之间的重要程度。这种综合损失策略有助于模型在训练过程中实现更全面的特征学习，使模型在实际的篡改检测中，能够更准确地识别边缘细节。Chen 等[26] 设计了一个多视图多监督的图像篡改检测定位框架，主要包括边缘监督分支和噪声敏感分支。在边缘监督分支中，为了更好地提取边缘特征信息，引入了 Sobel 层和边缘残差块。Sobel 层是一种滤波器，用于增强图像中的边缘特征。边缘残差块用于进一步处理从 Sobel 层得到的特征，它通过残差连接来增强对边缘特征的关注。此外也应用了多尺度的组合损失监督，分别是常规的像素尺度损失、针对边缘分支的边缘尺度损失以及图像级尺度损失。Salloum 等[51] 设计了一种基于全卷积网络的双分支网络来用于检测图像拼接篡改。该方法使用区域分支学习篡改区域的内部特征，利用边缘分支学习篡改边界处的特征，并提出了一种边缘增强的方法。首先，对区域分支和边缘分支得到的概率图进行阈值化处理，得到区域预测掩码图和边缘预测掩码图。接着对边缘预测掩码进行填充操作，以填充边缘预测掩码内部的空洞区域，获得增强后的边缘预测掩码。然后，将增强后的边缘预测掩码与区域预测掩码图融合，以获得最终的篡改定位掩码图。Rao 等[25] 设计了一种改进的基于条件随机场 (conditional random field, CRF)

的注意力模块 (improved CRF-based attention model, ICRF-Att), 其作用是生成边缘注意力图, 来突出显示篡改区域的边界。具体来说, 它基于 CRF 模型, 通过考虑像素之间的相互关系来优化篡改区域的识别, 其生成的注意力图能够在视觉效果上显著地突出篡改区域的边缘。在损失函数设计上, 与文献 [51] 的方法类似, 该方法也采用了边缘分支和区域分支相结合的监督机制, 边缘监督施加在生成的边缘注意力图上, 模型在 2 个损失函数的共同监督下进行训练, 从而优化其性能并完成最终的检测定位任务。Sun 等[52] 提出了一种包含边界引导模块和软边界监管策略的篡改检测方法。边界引导模块不仅可以挖掘篡改边界上的伪造痕迹, 而且可以引导模型关注边界内外周围的细微特征差异。软边界监管策略则通过滑动窗口的方式, 测量每个像素距离绝对边界[53] 的距离, 得到一个表示像素距离边界程度的软边界标签。这种软边界标签从本质上来说就是虚化了边缘的绝对范围, 将清晰的边缘变得更加地朦胧模糊。这种方法可以考虑不同像素的重要性, 提供更合理的边界表达。Ma 等[54] 设计了一种边界监督策略。首先, 运用数学形态学运算: 膨胀 (dilation) 和腐蚀 (erosion), 从原始的包含整个篡改区域的掩码图像生成突出边缘区域的边缘掩码图像。然后, 设计了特定的边缘监督损失函数来优化边缘特征的检测效果, 使得模型在预测图像篡改时更加关注边缘区域的变化。Li 等[27] 利用边缘信息作为引导, 提出一种用以解决伪造区域和真实区域存在特征耦合问题的方法。首先, 从 RGB 图像中提取了粗糙特征, 然后对粗糙特征应用 2 步处理。第 1 步, 通过边缘重建模块提取边缘特征。第 2 步, 应用图学习方法来处理粗糙特征, 以获得不同节点间的注意力关系。接着, 设计了一个区域消息传递控制器, 在边缘特征的引导下, 重新计算刚刚获得的注意力关系, 从而切断伪造区域与真实区域之间的联系, 解决了特征耦合问题, 实现了更精准的区域定位。实验结果表明, 该方法在多个数据集上的篡改检测性能优于其他方法。

1.4　基于对比学习的方法

对比学习是一种无监督学习技术, 其核心思想是通过比较不同样本之间的相似性或差异性来学习有效的特征, 如图 7 所示。通过这种方式, 可以将相似的样本映射到接近的空间区域, 从而使得相似的样本在特征空间中更加接近 (图中细箭头所示), 差异性较大的样本则在空间中彼此分开 (图中粗箭头所示)。在图像篡改检测定位任务中, 对比学习策略可以增强篡改区域与真实区域之间的特征差异性, 帮助模型学习到更具有区分度的篡改特征, 提高篡改检测的精度。

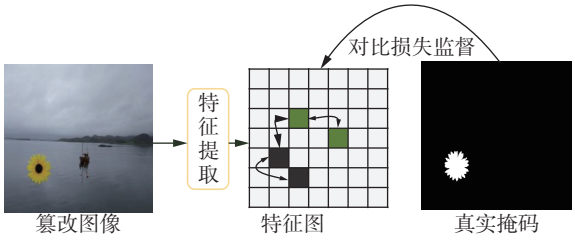

图 7　基于对比学习的方法

Fig. 7　Contrastive learning based methods

Niloy 等[43] 依据篡改区域和真实区域之间的特征差异性提出了一种基于对比学习的篡改检测方法。首先使用双流架构提取篡改图像中的篡改特征; 然后, 引入对比学习模块, 通过比较图像中不同区域的像素嵌入特征, 旨在使正常区域和伪造区域的特征分布得到明显分离。理想情况下, 模块中的对比损失计算是可以逐像素计算的, 即逐个计算每个像素点与真实掩码的损失。但实际上, 以这样的方式计算对比损失存在着 2 个主要的局限性: 首先, 基于逐个单像素的对比损失计算并没有考虑到相邻像素之间的关系; 其次, 逐个计算需要存储大小为 $HW\times HW$ 的矩阵, 这样的操作非常消耗内存。为了更有效地解决这 2 个问题, 该方法引入了一种创新的对比损失计算方法, 即将逐像素的计算方式转换为逐区域的计算方式。具体来说, 将特征图划分为大小相同的区域块, 然后计算每个区域块之间的对比损失。对于每个区域块, 对比损失函数评估了该区域块特征与同一图像中其他区域块特征之间的相似性。其计算公式为

$$L_i=\frac{1}{|A_i|}\sum_{k^+\in A_i}-\log\frac{\exp(\boldsymbol{f}_i\cdot\boldsymbol{k}^+/\tau)}{\exp(\boldsymbol{f}_i\cdot\boldsymbol{k}^+/\tau)+\sum\limits_{k^-}\exp(\boldsymbol{f}_i\cdot\boldsymbol{k}^-/\tau)}$$

式中: $\boldsymbol{f}_i$ 表示查询像素嵌入, $\boldsymbol{k}^+$ 是指与 $\boldsymbol{f}_i$ 具有相同标签的像素嵌入, 相对地, $\boldsymbol{k}^-$ 是指与 $\boldsymbol{f}_i$ 具有不同标签的像素嵌入, A_i 表示 $\boldsymbol{k}^+$ 像素嵌入的集合, τ 是温度超参数。所有的嵌入在使用损失函数之前都进行了 L2 归一化。这种方法使得相同标签嵌入像素更相似, 不同标签的像素块更排斥, 模型可以学习到更具判别性的特征表示, 从而提高图像篡改检测方法的定位性能。Zeng 等[55] 也提出了一个基于对比学习的篡改检测方法, 该方法使用 2 个流 (RGB 流和噪声流) 作为输入, 通过提取 2 个流的不同特征构建正负样本对, 与其他方法

相比，该方法基于目标检测框架，因此可以使用边界框注释进行训练，而不需要像素级的标注，因此计算资源更低。Hao 等[56] 提出了基于对比学习和边缘的双阶段通用检测框架，将整个任务分为 2 阶段，粗定位阶段与精细定位阶段。在粗定位阶段，分别利用边缘检测模块预测篡改区域的边缘分布，对比学习模块聚合正样本对的特征分布并区分负样本对的特征分布。在对比损失设计方面，作者提出篡改区域之间存在内在的相关性，但是这种相关性在真实区域中并不存在。真实区域的样本对会阻止对比学习获得更好的表示。 因此，只为被篡改的像素构建正对和负对。这样，模型可以学习到特征分布之间的差异，从而提高泛化能力。在精细定位阶段，模型通过多尺度融合模块聚合粗定位阶段的特征，生成精细的定位结果。Wu 等[57] 则将对比学习和无监督聚类结合到一起，提出了一种新颖、简单且有效的篡改检测方法。首先，在训练阶段，通过对比学习直接在像素级别对提取的特征进行监督。具体来说，使用改进的 InfoNCE 损失函数来计算对比损失，其公式为

$$\mathcal{L}_{\text{InfoNCE ++}} = -\log \frac{\dfrac{1}{J}\displaystyle\sum_{j\in[1,J]} \exp\left(\boldsymbol{q}\cdot\boldsymbol{k}_j^{+}/\tau\right)}{\displaystyle\sum_{i\in[1,K]} \exp\left(\boldsymbol{q}\cdot\boldsymbol{k}_i^{-}/\tau\right)}$$

式中：τ 是温度超参数，$\boldsymbol{q}$ 代表查询向量，$\boldsymbol{k}^{+}$ 是正样本的特征，$\boldsymbol{k}^{-}$ 是负样本的特征，J、K 分别代表正样本和负样本的数量。通过计算查询向量 $\boldsymbol{q}$ 与正负样本的相似度来训练模型，使得正样本在特征空间中更加接近查询向量，而负样本则更加远离。通过优化该损失函数，模型可以学到更有区分性的特征表示。此外，与传统的逐批次机制计算损失不同，该方法采用了逐个图像的损失计算方法，使得优化过程更加稳定并且可快速收敛。其次，在测试阶段，采用即时聚类算法 HDBSCAN(hierarchical density-based spatial clustering of applications with noise) 将提取的特征分别映射到最终的伪造掩码。通过结合对比学习和无监督聚类，能够显著地提升篡改检测定位的性能。Zhou 等[58] 提出了一种基于非互斥对比学习且无需预训练的篡改检测定位方法。该方法重新定义了对比学习的问题域，在原有篡改图像块和真实图像块的基础上，增加了同时包含篡改区域和真实区域的轮廓块，篡改块和真实块毋庸置疑是互斥的，而轮廓块对于它们则是非互斥的。为了更好地利用非互斥的轮廓块，提出了一种具有双分支的枢轴结构，其能够在训练时不断地在正负之间切换轮廓块的角色，并设计了与之匹配的枢轴损失，通过这种方式，解决了非互斥对比学习的配对问题。此外，该方法通过自监督对比学习从真实的篡改图像中生成大量的对比样本，从而无需大规模图片的预训练，能够有效解决训练数据不足的问题。

值得注意的是，以上基于边缘信息以及基于对比学习的方法从广义角度来说均为基于特征相似性的方法。除此之外，部分工作采用孪生网络结构的方式来构建特征相似性关系。孪生网络由 2 个结构相同且参数共享的子网络组成，它通常用于比较 2 个样本的相似程度，因此常常被应用于某些特定的篡改场景下，例如复制移动篡改检测等。Barni 等[59] 为了区分复制移动篡改中的源区域和目标区域，提出了一种双分支的孪生网络架构。他们认为，复制移动篡改中源区域和目标区域存在插值痕迹和边界不一致的问题，因此通过双分支的孪生网络架构，可以学习并比较它们之间的像素关系，从而区分源区域和目标区域。此外，还有一类特定的拼接篡改检测场景，称为约束条件下的拼接篡改检测问题。简而言之，提供一张待检测图片与一张供体图像，约束拼接检测的目的为判断待检测图片中是否存在有来自供体图像的区域。Liu 等[42] 提出了一种基于对抗学习的约束图像拼接检测与定位方法，该网络由 3 部分组成：DMAC 网络 (deep matching network based on atrous convolution)、检测网络、判别网络。DMAC 网络通过接收待检测图片和供体图片并生成准确的篡改概率图，而后由孪生网络架构的检测网络通过学习图像特征和篡改概率图之间的关系，来判断图像中的每个像素是否属于篡改区域。最后判别网络通过学习生成的篡改概率图和真实的篡改概率图之间的差异，来指导 DMAC 网络生成更准确的篡改概率图。实际上文中的 3 部分网络结构都在不同程度上借鉴了孪生网络的架构，通过这种架构可以更好地比较图像对的相似度，从而进行约束图像拼接检测定位。

1.5　基于 Transformer 等其他网络架构的方法

上文对于基于深度学习的图像篡改检测方法的共性关键技术进行了总结。以上共性关键技术对不同的深度学习网络架构具有一定的普适性，但基本都基于 CNN 展开。除 CNN 架构之外还有许多基于其他网络架构的方法，例如自编码器网络、LSTM(long short-term memory) 网络、生成对抗网络 (generative adversarial network, GAN) 和 Trans-

former 网络等。Zhang 等[60] 基于自编码器网络提出了一种双阶段的方法进行图像篡改检测, 首先利用堆叠自编码模型进行复杂的特征学习, 提取输入图像块的特征, 然后结合不同图像块的上下文信息进一步提高检测结果的准确性。Wu 等[31] 提出了一种基于 VGG(visual geometry group) 网络和 LSTM 网络的篡改检测定位方法, 该方法由基于 VGG 的特征提取器和基于 LSTM 的检测模块组成, 其中特征提取器针对 385 种图像处理类型进行训练以学习鲁棒的图像篡改痕迹, 而后经 LSTM 模块得到篡改区域的预测结果。Islam 等[61] 提出了一种基于生成对抗网络的双阶注意力模型, 用于图像复制移动伪造检测。在生成器网络中, 输入图像通过预训练的 VGG-19 网络的前 3 个块进行特征提取, 然后将提取的特征拼接起来, 利用提出的双阶注意力模块计算图像内部不同部分之间的关联, 生成区域注意力图和共现注意力图。同时通过 2 个不同的参数设置的空洞空间金字塔池化 (ASPP) 模块, 提取上下文特征, 并与注意力图进行元素级乘法, 以获取可能的复制移动区域的特征。这些特征被用于检测分支生成检测输出分数, 以及用于定位分支生成预测掩码。在判别器网络中的输入是图像和掩码的组合, 其任务是区分预测掩码是否与真实掩码相同, 通过对比预测掩码和真实掩码, 判别器评估生成掩码的质量, 从而判断生成的掩码是否准确地反映了图像中的实际篡改区域。判别器的反馈用于优化生成器, 使其生成的掩码越来越接近真实掩码, 从而提高生成模型的整体性能和准确性。通过这种对抗训练的方式, 生成器和判别器共同进步, 推动生成器生成更加真实和准确的掩码。Zhou 等[53] 提出了一种基于生成对抗网络的图像篡改分割方法, 该方法包括生成阶段、分割阶段和替换阶段。生成阶段使用生成对抗网络来模拟图像篡改的过程, 生成逼真复杂的篡改图像。在分割阶段, 分割网络依据篡改区域边缘的篡改痕迹特征来确定篡改区域及其边缘。在替换阶段, 将预测的篡改边界替换为原始真实图像区域, 并将新的篡改图像反馈给分割阶段作为新的训练样本。通过 3 个阶段的交替训练, 可以使分割网络学习到更强的篡改检测定位能力。

Transformer 为近年来较受欢迎的深度学习网络架构。它是一种基于自注意力机制的深度神经网络, 最开始被应用于自然语言处理 (natural language processing, NLP) 任务。而近年来, 基于 Transformer 架构的方法在计算机视觉领域也备受关注[62], 其中代表工作是 2020 年 Vision Transformer (VIT) 模型。Transformer 架构中最重要的就是自注意力机制, 自注意力机制可充分考虑到图像中其他所有图像块的信息, 从而有效地捕捉到中不同图像块之间的复杂关系。其原理是: 对于每个输入的小图像块, 可以通过乘以不同的权重矩阵生成 3 个向量: 查询向量 ($\boldsymbol{Q}$)、键向量 ($\boldsymbol{K}$) 和值向量 ($\boldsymbol{V}$)。对于每一个$\boldsymbol{Q}$向量, 计算它与所有其他$\boldsymbol{K}$向量的相似度, 得到注意力权重, 而后使用得到的注意力权重与所有的其他$\boldsymbol{V}$向量加权求和, 得到最终的输出向量, 其包含了所有其他输入图像块的信息, 如图 8 所示。

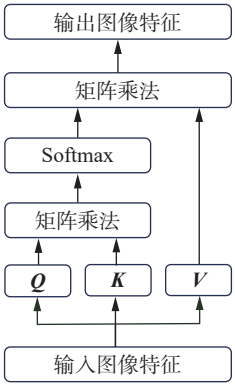

图 8　自注意力机制原理示意

Fig. 8　Schematic of self attention mechanism principle

注意力机制对应的计算公式为

$$A_{\text{ttention}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}$$

式中: d_k代表$\boldsymbol{K}$的维度, 它起到调节数值的作用, 防止$\boldsymbol{Q}$、$\boldsymbol{K}$计算的值过大, 从而使训练时的梯度保持稳定。

Transformer 的自注意力机制与传统 CNN 网络的卷积操作存在着很大的差异。传统的 CNN 网络主要依赖卷积核在局部区域内提取特征, 在更深的网络层数中可能会忽略全局依赖性, 即无法很好地捕获不同区域 (特别是非相邻区域) 之间的差异性。而拥有自注意力机制的 Transformer 网络正好可以弥补 CNN 网络的这部分缺陷。自注意力机制允许每个像素或图像块在计算时直接关注整个图像的所有其他像素或块, 这种全局依赖建模能力使得 Transformer 能够捕捉图像中的长距离关系, 而不是仅依赖于局部区域的特征。其次, Transformer 还引入了位置编码, 通过位置编码将空间信息融入到模型中, 提升对空间结构的理解, 从而有助于捕获篡改图像中的空间结构信息特征, 更好地理解图像中的篡改痕迹。

目前已有研究者提出了许多基于 Transformer 的图像篡改检测方法。Wang 等[37] 提出了

一种端到端多模态的图像篡改检测定位框架 ObjectFormer。ObjectFormer 包含高频特征提取模块、对象编码器和补丁解码器 3 个部分。首先，高频特征提取模块将图像从 RGB 域转换到频域，并提取多模态的补丁级别特征。然后，在对象编码器中，使用一组可学习的参数作为对象原型，并与刚刚提取的补丁级特征进行自注意力交互，使得对象原型可学习到图像中不同区域之间的依赖关系。补丁解码器则使用对象编码器更新后的对象原型来进一步细化补丁级特征，具体来说，补丁解码器首先对补丁嵌入和对象原型进行层归一化，然后将它们输入到注意力模块中以细化补丁嵌入。这样，每个补丁嵌入就可以进一步吸收来自更新后对象原型的信息。最后，通过边界敏感的上下文不一致性建模 (boundary-sensitive contextual incoherence modeling, BCIM) 模块来检测像素级别的不一致性，以进行细粒度的特征建模。总结来说，对象编码器负责学习图像中不同对象的一致性，而补丁解码器则利用这些学习到的对象信息来细化图像中每个补丁的特征表示，两者共同工作以提高篡改检测和定位的准确性。Hao 等[45] 受 Transformer 自注意力机制的启发，提出一种基于密集自注意力机制的图像伪造定位框架 TransForensics。该方法采用 ResNet-50 作为特征提取的主干网络，分阶段提取 5 个特征图。随后，将最后 4 个阶段的特征输入到标准的 Transformer 编码器网络中，以进一步学习和提取这些特征。在 Transformer 编码器网络中，密集自注意力机制被用来建模全局上下文以及不同尺度的图像补丁，从而提取细微的篡改痕迹特征。在预测阶段，密集校正模块会对各阶段的预测掩码进行重新校正，以获得更精确的检测结果。Liu 等[28] 提出了改进 VIT 的图像篡改检测定位网络 TBFormer(two-branch Transformer)，此方法设计了 2 个并行的 Transformer 编码器来分别提取 RGB 域和噪声域的特征信息。与 VIT 相同，该方法将提取的 RGB 图和噪声图分为 16 像素×16 像素的图像块输入到构建的特征提取器中。Lin 等[32] 提出了 ResNet 和 Transformer 相结合的篡改检测定位方法，使用传统的 ResNet 提取 RGB 图像的特征信息，采用 Swin-Transformer 从噪声图中提取全局和局部噪声特征。Ma 等[54] 提出了一种基于改进 VIT[63-65] 的图像篡改检测定位方法，称为 IML-VIT。他们认为保证篡改图像的高分辨率输入尤为重要，因此与目前大多数方法对输入图片进行调整到较小的尺寸不同，作者提出了保持高分辨率的填充方法，将图片填充到 1 024 像素×1 024 像素的尺寸，该策略能够保留每个图像的低级视觉信息，使模型能够更好地探索篡改特征。此外，为了平衡高分辨率带来的计算成本，使用窗口注意力模块替换掉 VIT 中的部分全局注意力块，在保证了全局信息传播的情况下，同时降低了复杂性。在与多个方法的对比下，该方法取得了良好的性能。Li 等[66] 提出了一种基于动态重要性感知的篡改检测框架。该框架通过引入动态变换器模块，将混合专家概念融入 Transformer 架构，使用多个特定的专家组来增强网络的泛化能力。特征重要性感知注意力机制自适应地感知不同区域的重要性，引导模型关注更具辨别性的区域，减少误报并减轻伪造与真实区域的特征耦合问题。Li 等[67] 提出了一种多视图 Transformer 框架，用于篡改检测和定位任务。框架通过多尺度监督的统一学习，整合 RGB、噪声和对象 3 个视图的信息。在特征编码阶段，使用双流架构处理 RGB 和噪声特征，并结合高频边缘特征，通过 BSFI-Net 提升判别能力。在特征对比协作阶段，RGB 和噪声特征融合后通过 Transformer Encoder 进行多尺度学习。最终，框架完成对象一致性建模、真实性二分类和篡改区域定位等任务。

虽然 Transformer 拥有众多优点，但其在图像篡改检测领域还存在许多挑战：首先，Transformer 模型复杂度较高且参数量庞大，极易在篡改数据集不足的情况下造成过拟合的问题，如 IML-VIT[54] 中，使用常见的 Xavier 方法或在 ImageNet-21k 对模型进行初始化时，模型完全不收敛或泛化性很差，最终作者采用 MAE(masked auto encoder) 初始化方法缓解了这个问题。其次，Transformer 的自注意力机制虽在全局信息建模上效果突出，但其局部的感知能力有限，可能无法很好地捕获篡改图像中局部的篡改痕迹，因此许多研究者考虑采用移位窗口改进的自注意力机制或通过 CNN 与 Transformer 相结合的方法来更全面地考虑局部和全局细节特征[32]。最后，在实际部署中，Transformer 等方法的模型复杂度较高且参数量庞大，这可能导致对计算资源的需求增加。以 IML-VIT 为例，当 Batchsize 设置为 1 时，显存占用已超过 12GB。因此，基于 Transformer 的网络模型在训练和部署环节对显存可能具有较高的需求。

综上所述，尽管 Transformer 在图像篡改检测任务中展现了不错的能力，但也面临着模型复杂度过高、易过拟合、局部信息建模能力有限、计算

资源需求增加等挑战, 仍需要进一步的研究和优化, 从而提升 Transformer 在图像篡改检测任务中的实际应用效果。

1.6　基于有损后处理背景下的方法

前几节介绍的各类方法都可以在图像未经过有损后处理的情况下取得理想的检测定位效果, 但在现实场景中, 篡改图像大多通过社交网络进行传播。在传播的过程中, 不可避免地会遭受一些有损处理, 比如 JPEG 压缩、随机噪声干扰等。这些操作可能会破坏篡改图像内部的篡改痕迹, 从而导致篡改检测定位的效果变差, 这给图像篡改检测定位任务提出了更严峻的挑战。

在日常生活中, 图片的压缩方式有很多, 而最常见的压缩方法为 JPEG 压缩, 因此目前大多数针对有损压缩的篡改检测方法都选择研究在 JPEG 压缩影响下提升篡改检测方法的定位性能。Rao 等[25] 提出了一种基于自监督域适应网络的方法, 该方法借鉴了迁移学习领域的领域自适应策略[68]。作者利用无压缩图像 (源域) 的篡改检测经验, 来对 JPEG 压缩图像 (目标域) 进行篡改检测。具体而言, 自监督域适应网络由 2 个主要组成部分构成: 孪生架构的骨干网络和压缩近似网络。压缩近似网络负责生成 JPEG 代理图, 这些代理图用于展现 JPEG 图像的普遍特性。随后, 孪生架构的骨干网络通过域适应策略, 学习从无压缩图像中获得的篡改检测经验, 并将这些经验有效地应用到 JPEG 压缩图像中。在训练过程中, 模型接受域损失、边缘损失和区域损失这 3 种损失函数的联合监督, 以确保模型能够准确地生成最终的预测结果。Kwon 等[39] 设计了一个压缩伪影追踪网络, 通过利用 DCT 系数来学习 JPEG 压缩的篡改痕迹。传统的 CNN 网络无法直接学习 DCT 系数的分布, 因为卷积会丢失对 DCT 系数至关重要的空间信息, 而该方法设计了一个 JPEG 伪影学习模块, 可以实现在不丢失空间信息的情况下学习 DCT 系数的分布。在大量基准数据集上的实验结果表明, 无论是针对 JPEG 压缩图像还是未压缩图像还是双 JPEG 压缩图像, 其都具有良好的篡改检测性能。Zhuang 等[69] 提出了一种新颖的图像篡改检测定位恢复方法。该方法包括图像恢复框架和篡改定位框架 2 部分, 核心思想是利用图像恢复框架恢复被扭曲的篡改图像, 生成篡改图像的高质量副本, 从而减弱 JPEG 压缩等操作对篡改图像的影响, 便于篡改定位框架识别篡改区域。由于图像恢复框架和篡改定位框架是 2 个相辅相成的组合网络, 因此在训练策略上, 作者采用了交替训练的方法进行优化恢复和定位框架, 具体来说, 在一个 epoch 内优化恢复框架, 而在下一个 epoch 内优化定位框架, 这样的策略使得定位框架中的损失下降得更快, 并且收敛得更早, 从而使得模型能够在更短的时间内学习到更为精细的篡改特征。值得注意的是, 该框架中所提出图像恢复模块是可迁移的, 即当直接和另一个篡改定位模块一起进行部署时仍然有效。文中作者采用了 3 种不同的篡改定位方法模块进行实验, 通过实验表明, 与未使用图像恢复框架的篡改定位方法相比, 添加图像恢复框架的篡改定位方法均取得了更优秀的效果。

上述方法都是将 JPEG 压缩处理作为代表来评估算法的对后处理操作的鲁棒性, 但篡改图像经在线社交网络 (online social networks, OSNs) 传播的过程中会经历除 JPEG 压缩之外的其他许多有损操作。因此, 上述方法仍无法很好地对网络中传播的篡改图像案例实现精准的检测。Wu 等[70] 则提出了一种新颖的训练方案, 旨在提高模型对 OSN 后处理图像的鲁棒性。由于在 OSN 传输图像的过程中, 会引入很多的噪声, 篡改检测模型需要了解这些噪声从而提升对噪声的鲁棒性。但是目前各种社交平台并未公开在传输过程中对图像的处理流程, 并且不同的社交平台对图像的处理方法也不同。因此, 作者首先考虑对 OSN 引入的噪声进行彻底的分析, 将引入的噪声分为 2 部分, 即可预测噪声和未知噪声, 并分别对其建模。可预测噪声主要是模拟一些已知操作带来的损失, 比如调整图片大小、JPEG 压缩等。对于这些噪声, 作者采用一个改进的 Unet 结构和一个可微的 JPEG 层来模拟 OSN 的操作。未知噪声是对可预测噪声的补充, 主要模拟预测 OSN 平台的未知操作。显然, 从信号本身的特征角度来建模未知噪声是不现实的, 于是作者采用一种新的思路, 即利用对抗噪声的思想来对未知噪声建模, 将关注点转移到检测器本身, 只关注那些可能降低检测性能的噪声, 而忽略那些对检测影响不大的噪声。在完成对已知和未知的噪声建模后, 作者采用一个改进的 SE-U-net 模型来训练学习获得的噪声特征, 从而提升模型对 OSN 图像检测的鲁棒性。在多个基准数据集上进行实验发现, 在对多个主流社交网络平台 (微信、微博、Facebook) 传播的篡改图像检测中, SE-U-net 获得了优秀的效果表现。

基于上述总结, 本文进行如下对比分析: 从共性关键技术的角度来说, 多流信息方法以及多尺

度信息方法均为采用特征融合的方式，通过融合不同输入域以及不同尺度信息的方法提升了篡改细节痕迹的识别精度。基于边缘信息以及基于对比学习的方法均基于特征相似性展开，从而达到了增强真实区域以及篡改区域的特征差异性的目的。从神经网络基础架构来说，近年来卷积神经网络仍为图像篡改区域定位的主流架构，GAN 以及 Transformer 等架构作为补充。在未来，Transformer 是否可以在图像篡改检测领域全面地超越 CNN，仍具有较大的探索空间。从应用场景的角度来说，面向有损压缩等现实场景的方法侧重于提升图像篡改检测方法的鲁棒性，具有较高的应用意义以及研究价值。

2　数据集与评价指标

2.1　数据集

图像篡改检测数据集一般包含原始图像和篡改图像，以及与篡改图像相对应的像素级真实掩码 (ground-truth, GT)。本节汇总了目前领域内广泛使用的篡改数据集，如表 1 所示。下面将详细的介绍每个数据集的内容和特点。

表 1　常用数据集汇总

Table 1　Summary of commonly used datasets

数据集名称	发布时间	篡改方式	图片数(真/假)	图像尺寸/像素×像素	图片格式	GT
Columbia[71]	2006	拼接	183/180	757×568 ~ 1 152×768	TIF、BMP	有
MICC F220[72]	2011	复制移动	110/110	722×480 ~ 800×600	JPG	无
MICC F2000[72]	2011	复制移动	1 300/700	2 048×1 536	JPG	无
CASIA v1[73]	2013	拼接、复制移动	800/921	374×256	JPG	有
CASIA v2[73]	2013	拼接、复制移动	7 200/5 123	320×240 ~ 800×600	JPG、BMP、TIF	有
DSO-1[74]	2013	拼接	100/100	2 048×1 536	PNG	有
CoMoFoD[75]	2013	复制移动	260/260	512×512 ~ 3 000×2 000	PNG、JPG	有
CMH[76]	2015	复制移动	—/108	845×634 ~ 1 296×972	PNG、JPG	有
GRIP[77]	2015	复制移动	80/80	1 024×768	PNG	有
Coverage[78]	2016	复制移动	100/100	400×486	TIF	有
Wild Web[79]	2015	现实实例	90/9 657	72×45 ~ 3 000×2 222	PNG、BMP、JPG、GIF	有
RTD-Korus[80]	2016	拼接、复制移动	220/220	1 920×1 080	TIF	有
NIST 16[81]	2016	拼接、复制移动、删除	560/564	500×500 ~ 5 616×3 744	JPG	有
NIST 17[81]	2017	多种操作	2 667/1 410	160×120 ~ 8 000×5 320	RAW、PNG、BMP、JPG	有
PS-Battles[82]	2018	多种操作	11 142/102 028	130×60 ~ 10 000×8 558	PNG、JPG	无
MFC2018[81]	2018	多种操作	14 156/3 265	128×104 ~ 7 953×5 304	RAW、PNG、BMP、JPG、TIF	有
MFC2019[81]	2019	多种操作	10 279/5 750	160×120 ~ 2 624×19 680	RAW、PNG、BMP、JPG、TIF	有
DEFACTO[83]	2019	多种操作	—/229 000	240×320 ~ 640×640	TIF	有
IMD2020[84]	2020	多种操作	37 010/37 010	193×260 ~ 4 437×2 958	PNG、JPG	有

注：表中“—”表示并未给出具体数据。

Columbia 数据集[71] 于 2006 年发布，它由 183 幅真实图像和 180 幅篡改图像组成，这些图像都是彩色图像且包含 Ground-truth。但是该数据集仍存在着一些主要的缺陷：拼接区域不进行任何后处理操作，篡改痕迹清晰可见，导致篡改图像并不真实，远不能够和现实世界中的篡改图像相比，且该数据集仅包含未被压缩的图像格式。

MICC 数据集[72] 于 2011 年发布，是可用数据集中发布时间最早和最常用的数据集之一，主要针对复制移动伪造检测。它由 4 个子集组成，但最常用的还是 MICC F220 和 MICC F2000 这 2 个子集。为了使篡改图更符合现实中的篡改实例，制作者通过随机选择图像中的矩形区域，然后对其应用旋转或缩放类型的操作来获得篡改图像。但是该数据集依旧存在 2 个缺点：第一，仅仅使用了旋转和缩放 2 种类型的操作来获得篡改图像，并没有进行例如添加噪声和应用 JPEG 压缩等后处理操作；第二，数据集并没有提供相对应的 Ground-truth，而 Ground-truth 是辅助分类器进行训练的重要依据之一。由于缺乏 Ground-truth，导

致最终对检测精度的评估时，依据的不是篡改图像中实际的篡改区域，而是可以被检测为篡改图像的数量。这使得使用该数据集时的评估并不准确，也无法反映模型算法的实际性能。

CASIA 数据集[73]是图像篡改检测领域中最常用的数据集之一。CASIA 数据集通常被分为 CASIA v1 和 CASIA v2 2 个子数据集：在 CASIA v1 中，一共包括 1 721 张图像，其中 800 张真实图像，921 张篡改图像，图像的尺寸都限制在 374 像素×256 像素。在 CASIA v1 中，篡改边缘的痕迹比较明显，比较容易被检测到。在 CASIA v2 中，图像会经过一定程度的后处理操作，使其在拼接复制移动篡改后更好地适应真实场景。CASIA v1 主要通常用于评估和比较不同算法在检测和分析图像篡改方面的性能。CASIA v2 一般用于训练阶段。

DSO-1 数据集[74]是来自电气电子工程师学会 (institute of electrical and electronics engineers, IEEE) 图像取证挑战赛数据集的一个子集，它主要针对图像拼接篡改方式。它由 200 张图片组成，篡改图像和真实图像分别有 100 张。虽然数量较少，但其篡改的质量很高，大多数图像都符合现实生活中的篡改实例，并且提供相对应的 Ground-truth。

CoMoFod[75]是针对复制移动篡改方式常用的数据集，由 Tralic 等于 2013 年发布，它包含 260 张篡改图像，并且提供 Ground-truth。与之前提到的数据集最大的不同点在于：此数据集包含更多的攻击处理，如旋转、缩放、噪声处理、JPEG 压缩、图像模糊、亮度变化、对比度调整等，这能够在一定程度提高篡改检测模型的鲁棒性。

CMH 数据集[76]同样是针对复制移动篡改方式的数据集。共有 108 张复制移动伪造图像，提供 Ground-truth。篡改图像包含旋转和缩放操作变化的攻击。缺点是数据集包含的图像较少，可能不足以准确地评估复制移动算法的性能。

GRIP[77]由 Cozzolino 等构建，包含 80 张真实图像和 80 张篡改图像，并且提供 Ground-truth。此数据集的优点是复制移动区域的大小变化较小，缺点是缺乏常见的攻击处理操作。

Coverage[78]是针对复制移动篡改方式目前最常用的数据集之一，包含 100 张篡改图像，并且提供 Ground-truth。该数据集的图像经过发布者精心挑选，具有多个相似的真实的物体，这使得篡改变得更加逼真，同时也给篡改检测提出了更高程度的挑战。缺点是图像数量有限，并且缺乏一些后处理攻击操作，一般用于评估阶段。

Wild Web 数据集[79]与其他合成数据集有所不同，它是基于互联网中真实篡改的案例而构建的，一共有 9 000 多张篡改图像，作者付出了很大的努力收集相同图像的不同篡改版本并且提供了对应的 Ground-truth。

RTD-Krous 数据集[80]是由 Krous 等提出的，包含着各种类型的篡改图像，共 220 张真实图像和对应的篡改图像，提供 Ground-truth，且数据集的图像非常逼真。

NIST 16[81]是由美国国家标准与技术研究院发布数据集。共包含 1 124 张图像，其中 564 张篡改图像。NIST 16 数据集中的每个视觉上看起来内容相同的篡改图像都包含 4 类不同的版本：高质量和低质量的 QF(quality factor) 压缩版本，篡改区域的边界是否经过后处理操作的版本。随后，该机构又陆续发布了 NIST 17、MFC2018、MFC2019。与 NIST16 不同，后续的 3 个数据集中相同的篡改图像不再包含多种不同的版本，但篡改图像的数量明显增加，篡改的方式也多种多样。

虽然上述提到的数据集的样本数量最多达到了 1 万多张，但是这对基于深度学习的模型训练来说是远远不够的。为了解决这个问题，研究者们开始着力于创建大规模大容量的数据集。PS-Battles 数据集[82]是一个包含超过 10 万图像的数据集，包含原始的图像和与其对应的数量若干的不同版本的篡改图像。DEFACTO 数据集[83]是由 Mahfoudi 等在 MSCOCO 数据集的基础上进行构造的，它包含接近 23 万张篡改图像，包含拼接、复制移动、删除、面部变形等多种篡改方式。并且所有的篡改图像都提供了对应的 Ground-truth。IMD2020 数据集[84]于 2020 年发布，它一共包含 37 000 多张篡改图片，其中 2 000 张来自于现实生活中真实篡改的案例，并提供对应的 Ground-truth。

通过总结近几年基于深度学习模型的篡改检测定位方法，本文注意到，目前基于深度学习的图像篡改检测方法仍然面临着训练数据集不足的问题[31,42,44,85]，虽然已存在部分大容量的数据集，但在实际实验过程中，应用得却并不广泛，不少篡改检测方法仍倾向于自主合成更大规模的数据集供实验使用。目前主流的合成方法是通过 MS-COCO 数据集[86]、SUN 数据集[87]、Dresden 数据集[88]作为源样本，再通过计算机脚本程序，进行随机裁剪、旋转等操作合成数据集，本文总结了部分具有代表性方法的自建合成数据集，如表 2 所示。

表 2　部分合成数据集汇总

Table 2　Partial composite dataset summary

方法名称	篡改方式	图片数(真/假)	源样本	后处理操作	GT
BusterNet[85]	复制移动	—/131 778	SUN、MSCOCO	—	有
ManTra-Net[31]	多种操作	—	Dresden	—	—
PSCC-Net[49]	拼接、复制移动、删除	81 910/294 829	MSCOCO、KCMI、VISION、Dresden	—	有
GCA-Net[33]	拼接、复制移动、删除	—/170 000	Dresden、MSCOCO、DeFACTO、IMD-Real	—	有
ObjectFormer[37]	拼接、复制移动、删除	—/62 000	MSCOCO、Paris StreetView	JPEG压缩、随机噪声	有
TBFormer[28]	拼接、复制移动、删除	—/156 006	CASIA v2、ADE20k	—	有
SAFL-Net[52]	拼接、复制移动、删除	21 301/34 000	MSCOCO、PSBattles	—	有
ERMPC[27]	拼接、复制移动、删除	—/64 000	MSCOCO	JPEG压缩、随机噪声	有
CAT-Net[39]	拼接、复制移动	—/800 000	MSCOCO、RAISE	JPEG压缩	有

注：表中“—”表示未给出具体数据。

2.2　评估指标

评估指标是用于描述模型和算法在基准数据集上的性能工具，图像篡改检测定位任务本质上是像素级别的二分类问题。因此，目前领域内广泛使用的评估指标包含：精确率 (precision)、召回率 (recall)、F1 分数 (F1-score)、ROC 曲线下面积 (area under the curve, AUC)、交并比 (intersection over uunion, IoU)。

在介绍上述评估指标之前，首先需要理解混淆矩阵的概念，混淆矩阵由 4 个部分组成：

TP(true positive)，被正确预测为伪造像素的伪造像素。

TN(true negative)，被正确预测为真实像素的真实像素。

FP(false positive)，被错误地预测为伪造像素的真实像素。

FN(false negative)，被错误地预测为真实像素的伪造像素。

2.2.1　F1-score

F1-score 是综合考虑精确率和召回率的评价指标，常用于衡量分类算法的性能，可以通过以下公式计算：

$$F_1 = 2 \times \frac{P_{\text{recision}} \times R_{\text{ecall}}}{P_{\text{recision}} + R_{\text{ecall}}}$$

式中：P_{recision}为精确率，它表示预测结果为篡改的样本占实际篡改样本的比例；R_{ecall} 称为召回率，它表示所有被篡改样本中被正确识别为篡改样本的比例。精确率和召回率的计算公式分别为

$$P_{\text{recision}} = \frac{T_{\text{P}}}{T_{\text{P}} + F_{\text{P}}}$$

$$R_{\text{ecall}} = \frac{T_{\text{P}}}{T_{\text{P}} + F_{\text{N}}}$$

F1-score 的取值范围为 0 ~ 1，数值越接近 1 表示算法的性能越好。F1-score 能够综合考虑算法的准确性和全面性，对于不平衡的数据集和任务具有较好的表现。

2.2.2　AUC

AUC 是一种常用的评价指标，通常用于衡量分类模型在二分类问题中的性能。ROC 曲线反映了不同阈值下真阳性率 (true positive rate, TPR) 和假阳性率 (false positive rate, FPR) 之间的权衡关系。ROC 曲线的横轴是假阳性率，即将负样本错误地预测为正样本的比例；纵轴是真阳性率，即将正样本正确地预测为正样本的比例。AUC 的值是在 ROC 曲线下的面积，取值范围为 0 ~ 1。当取值等于 0.5 时，表示模型的性能与随机猜测没有区别，即无法区分正样本和负样本，当取值越接近 1 时，表示模型的性能越好。在目前的研究工作中，一般将 F1-score 和 AUC 结合使用，以获得更全面的性能评估。

2.2.3　IoU

IoU 是一种常用的图像分割评价指标，用于衡量目标检测或图像分割算法的准确性。在图像篡改检测定位领域中，也常常被用于评估篡改定位性能的好坏。IoU 的计算方法是通过计算预测篡改区域与真实篡改区域的交集面积除以它们的并集面积，计算公式为

$$I_{\text{oU}} = \frac{T_{\text{P}}}{T_{\text{P}} + F_{\text{P}} + F_{\text{N}}}$$

IoU 的取值范围为 0 ~ 1，其中 0 表示没有重叠，1 表示完全重叠。通常情况下，IoU 越接近 1，表示预测的篡改区域与真实的篡改区域的重合度越高，模型的性能越好。

2.3　不同方法的性能对比

2.3.1　常规方法性能对比

为了对第 1 节中提到的方法进行直观的性能对比，本文对近几年具有代表性的方法进行了整理，如表 3 所示，这些方法大都发表在计算机视觉领域的顶级期刊会议上。对于评估数据集，本文选择最常用的 NIST16、Coverage、CASIA v1 作为代表。在评估指标的选择上，本文选择 F1-Score、AUC 和 IoU。表中的数据均来自各文献中原作者所提供的数据。

表 3　不同方法的性能对比

Table 3　Performance comparison of different methods

方法名称	应用策略	评估数据集								
		NIST16			Coverage			CASIA v1		
		F1-Score	AUC	IoU	F1-Score	AUC	IoU	F1-Score	AUC	IoU
RGB-N[22]*	M	0.722	0.937	—	0.437	0.817	—	0.408	0.795	—
MFCN[51]	E	0.571	—	—	—	—	—	0.541	—	—
ManTra-Net[31]*	M	—	0.795	—	—	0.819	—	—	0.817	—
TransForensics[45]	S+T	—	—	—	0.674	0.883	—	0.627	0.837	—
DFCN[89]*	S	0.310	0.750	0.230	—	—	—	—	—	—
MVSS-Net[26]	M+S+E	0.292	—	—	0.453	—	—	0.452	—	—
Rao+ComNet[25]	M+E+J	0.416	—	—	0.516	—	—	0.275	—	—
PSCC-Net[49]*	S	0.742	0.991	—	0.723	0.941	—	0.554	0.875	—
GCA-Net[33]*	M+S	0.845	0.953	—	0.695	0.874	—	—	—	—
SCSE-Unet[70]*	S+J	0.332	0.783	0.255	—	—	—	—	—	—
CAT-Net[39]*	M+S+J	0.556	—	—	0.413	—	—	—	—	—
ObjectFormer[37]*	M+T	0.824	0.996	—	0.758	0.957	—	0.579	0.882	—
TBFormer[28]*	M+S+T	0.834	0.997	—	—	—	—	0.696	0.955	—
EMT-Net[32]	M+S+T+E	0.825	0.987	—	0.353	0.812	—	0.459	0.856	—
SAFL-Net[52]*	S+E+C	0.879	0.997	—	0.803	0.970	—	0.740	0.908	—
CFL-Net[43]	M+S	—	0.997	—	—	—	—	—	0.863	—
ERMPC[27]*	M+S+E	0.836	0.997	—	0.773	0.984	—	0.586	0.904	—
TruFor[35]*	M+S+T	0.399	0.760	—	0.600	0.770	—	0.737	0.916	—
HiFi-Net[38]*	M+S	0.850	0.989	—	0.801	0.961	—	0.885	0.616	—
NCL[58]	S+C	0.831	0.912	—	0.801	0.928	—	0.598	0.864	—
IML-VIT[54]	S+E+T	0.339	—	—	0.425	—	—	0.658	—	—

注：表中“M”表示使用多流信息融合策略，“S”表示使用多尺度特征融合策略，“E”表示使用边缘信息策略，“C”表示使用对比学习策略，“T”表示使用Transformer架构，“J”表示应用有损后处理策略，“*”表示使用大规模合成训练集进行训练，“—”表示文献中并未给出相关的数据。

由表 3 数据本文可以总结出以下几点结论：

1) 融合多种关键技术能够在一定程度上提升篡改检测精度。如表 3 中显示的 ERMPC[27]、TB-Former[28]、TruFor[35]、HiFi-Net[38]、SAFL-Net[52] 等使用多种关键技术的方法精度均在一定程度上超过了早期的使用单独一项策略的部分方法，如 RGB-N[22]、ManTra-Net[31]、MFCN[51] 等。因此可以通过融合关键技术的方法在一定程度上提升篡改检测的精度。

2) 使用大规模训练数据集一定程度上可以提升方法的检测精度。表 3 中显示较多的方法采用了大规模的合成数据集进行了训练，如文献 [27,31,38-39,49] 等方法，在不同的测试数据集上均取得较高的检测精度。大量合成的数据集提供了更多样化的样本，并且合成数据集通常涵盖了多种真实篡改场景的变种，从而能够提高模型在面对真实篡改场景时的鲁棒性与泛化性。

3) 大部分篡改检测方法对于不同测试数据集的泛化性较差。由表 3 可以看出，目前的篡改检测方法很难在不同的数据集上同时达到较高的检测精度，如 EMT-Net[32] 在 NIST 数据集上的检测效果较好，但在 Coverage 数据集上效果却不是很理想。这是由于不同的数据集[73,78,81] 包含的篡改类型、生成篡改图的方法、压缩的质量，甚至篡改

区域的大小都可能存在不同，而目前的方法只能对训练数据集中的篡改方式达到较好的检测结果，较难兼顾所有的篡改特性，从而导致其泛化性较差。

4)Transformer 在图像篡改检测中还存在着一定的挑战。虽然 Transformer 在多种任务下都取得了显著的提升从而吸引了大量研究者的关注[62]，但表 3 中显示的基于 Transformer 的篡改检测方法 (例如 ObjectFormer[37]、TransForensics[45]、IML-VIT[54] 等) 相对于其他方法尚未取得识别精度以及泛化性上的显著领先，因此 Transformer 在图像篡改检测方面的应用尚处于探索阶段，仍需要进一步研究和优化以提升识别精度和泛化能力。

2.3.2　基于有损后处理背景的方法性能对比

由于基于有损后处理背景下的检测定位方法相对较少，因此本文仅总结了几个具有代表性的方法，如表 4、5 所示。表 4 涉及了在不同 JPEG 压缩因子下的性能对比，表 5 则总结了在不同社交平台传输后的篡改图片的检测性能对比。在不同的压缩因子和社交平台传输条件下的表现来看，相较于标准数据集方法，经过精心设计的有损后处理检测方法在大多数场景中均取得了最佳成绩，这表明该方法在面对现实场景时，具备更强的鲁棒性和更高的定位精度。此外，可视化对比如图 9 所示，CAT-Net 在处理 NIST16 数据集的删除样本时，能够精确定位细微的篡改区域；而 OSN 在处理 NIST16 数据集的拼接样本时，也实现了最佳的检测定位效果。这进一步表明，在极端现实场景下，这类方法有更优的性能表现。

表 4　针对 JPEG 压缩方法的性能对比

Table 4　Performance comparison of JPEG compression methods

数据集	方法	策略	压缩因子								
			QF60			QF70			QF80		
			F1-Score	AUC	IoU	F1-Score	AUC	IoU	F1-Score	AUC	IoU
DEFACTO	DFCN[89]	—	0.402	0.889	0.338	0.409	0.890	0.345	0.463	0.907	0.398
		R	**0.421**	**0.896**	**0.354**	**0.432**	**0.896**	**0.365**	**0.485**	**0.910**	**0.415**
	SCSE-Unet[70]	—	0.588	**0.951**	0.518	0.594	**0.953**	0.525	0.629	**0.959**	0.562
		R	**0.600**	0.950	**0.533**	**0.604**	0.952	**0.538**	**0.634**	0.958	**0.567**
	MVSS-Net[26]	—	0.461	**0.914**	0.392	0.467	**0.916**	0.398	0.507	0.900	0.432
		R	**0.477**	0.910	**0.404**	**0.489**	0.911	**0.415**	**0.532**	**0.925**	**0.455**
IMD2020	DFCN[89]	—	0.287	0.829	0.186	0.289	0.835	0.188	0.290	0.836	0.189
		R	**0.338**	**0.841**	**0.229**	**0.344**	**0.854**	**0.233**	**0.357**	**0.861**	**0.244**
	SCSE-Unet[70]	—	0.429	**0.897**	0.320	0.429	0.900	0.321	0.436	0.903	**0.356**
		R	**0.446**	**0.897**	**0.334**	**0.454**	**0.901**	**0.340**	**0.459**	**0.905**	0.347
	MVSS-Net[26]	—	0.320	**0.813**	0.224	0.331	**0.814**	**0.234**	0.323	**0.813**	**0.231**
		R	**0.331**	0.812	**0.230**	**0.334**	**0.814**	0.233	**0.325**	0.812	0.228

注：表中“—”表示使用正常训练策略，“R”表示使用针对JPEG的图像恢复框架，最好的结果使用加粗处理，DEFACTO是完整数据集的一个子集。

表 5　不同 OSN 下的方法性能对比

Table 5　Comparison of method performance under different OSNs

社交平台	方法名称	评估数据集								
		NIST16			CASIA v1			Columbia		
		F1-Score	AUC	IoU	F1-Score	AUC	IoU	F1-Score	AUC	IoU
—	ManTra-Net[31]	0.088	0.634	0.054	0.130	0.776	0.086	0.357	0.747	0.258
	DFCN[89]	0.250	0.778	0.204	0.192	0.654	0.119	0.541	0.789	0.395
	SCSE-Unet[70]	**0.332**	**0.783**	**0.255**	**0.509**	**0.873**	**0.465**	**0.707**	**0.862**	**0.608**
Facebook	ManTra-Net[31]	0.095	0.652	0.057	0.102	0.763	0.065	0.103	0.626	0.103
	DFCN[89]	0.207	0.705	0.138	0.190	0.654	0.116	0.479	0.687	0.338
	SCSE-Unet[70]	**0.329**	**0.783**	**0.253**	**0.462**	**0.862**	**0.417**	**0.714**	**0.883**	**0.611**

续表 5

社交平台	方法名称	评估数据集								
		NIST16			CASIA v1			Columbia		
		F1-Score	AUC	IoU	F1-Score	AUC	IoU	F1-Score	AUC	IoU
Weibo	ManTra-Net[31]	0.088	0.671	0.053	0.099	0.754	0.063	0.103	0.620	0.056
	DFCN[89]	0.192	0.706	0.125	0.191	0.653	0.117	0.458	0.676	0.319
	SCSE-Unet[70]	**0.294**	**0.780**	**0.219**	**0.466**	**0.858**	**0.421**	**0.724**	**0.883**	**0.626**
Wechat	ManTra-Net[31]	0.095	0.654	0.057	0.080	0.724	0.048	0.199	0.613	0.125
	DFCN[89]	0.176	0.701	0.114	0.193	0.651	0.119	0.487	0.676	0.344
	SCSE-Unet[70]	**0.286**	**0.764**	**0.214**	**0.405**	**0.833**	**0.358**	**0.727**	**0.883**	**0.631**

注：表中“—”表示篡改数据集未经社交平台传输，最好的结果使用加粗处理。

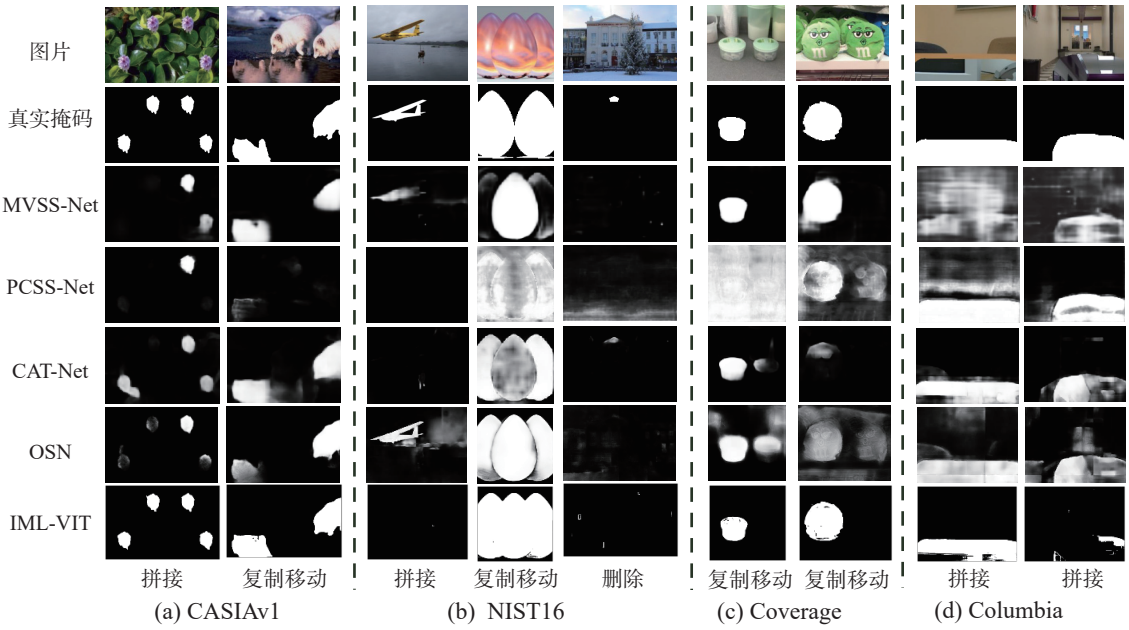

图 9　篡改区域检测结果可视化对比

Fig. 9　Visual comparison of tampering area detection results

2.3.3　可视化对比

为了能够更加直观地展示各种方法的性能，本文对篡改检测结果进行了可视化。从 4 个最常用的评估数据集 (CASIAv1、NIST16、Coverage、Columbia) 中选择可视化样本，对于每个数据集，最少选择 2 张图片进行对比，并保证选择的图片能够覆盖数据集中全部的篡改类型。为了公平起见，本文仅采用了目前开源代码并提供预训练模型的方法进行测试，如图 9 所示。通过可视化对比，可以观察到大多数的篡改检测方法能够较精确地定位篡改检测区域，但在某些方面仍存在不足。例如，大部分方法在面对 NIST16 数据集上的微小篡改区域检测方面仍存在较大困难。

3　问题与展望

3.1　问题

尽管基于深度学习技术的图像篡改检测方法取得了显著的成果和良好的检测效果，但在越来越先进的篡改技术和现实环境的挑战下，仍然存在许多问题[31,42,58,70,85]，本文将其主要概括为以下几点。

1) 大规模训练数据集的缺乏：缺乏大规模权威的训练数据集是图像篡改检测的主要问题之一。对于篡改检测而言，高水平数据集的制作非常困难。一方面，数据集需要包含不同大小、格式、篡改方式、压缩级别的图片，并且还需要保证篡改图像的“真实性”，即篡改图像要符合现实生活中真实的篡改案例，能够对人的认知造成迷惑。另一方面，生成数以万计“真实”的篡改图像也绝非易事，如在 2.1 节所述，很多研究者通过 MSCOCO 等数据集通过随机裁剪、翻转等方法进行自主合成大规模的训练数据集，这在一定程度上解决了训练所需大规模数据量的问题，但是通过这种方式生成的篡改数据集质量参差不齐，大

部分的篡改图像仅用肉眼就可以看出不同，与实际的篡改案例有很大的差距。而在这种具有明显偏差上的数据集进行模型的训练，可能会导致最终检测定位的性能的降低。此外，使用合成数据集训练还会导致各类篡改检测方法无法进行公平的性能对比。由此可见，创建一个大规模权威的训练数据集是图像篡改检测任务急需解决的问题之一。

2) 跨数据集的泛化性差：目前许多篡改检测方法都存在着训练模型在一个数据集上表现较好，但在其他数据集上表现较差的情况。这表明提出的篡改方法无法应对现实世界复杂和多样化的篡改实例，因此需要研究者在不同的角度进行深入的研究，提高检测方法的泛化性。

3) 对有损后处理的鲁棒性差：现实场景中，篡改图像大多通过社交网络进行传播。传播的过程中，必然会受到一些有损后处理，例如 JPEG 压缩等。这种有损后处理操作会严重破坏图像内部的篡改痕迹，导致目前篡改检测方法的性能下降。而目前针对有损后处理的研究工作也相对较少，很难完全的应对实际的篡改场景。因此，如何提高检测模型的鲁棒性也是一个具有挑战性的问题。

3.2　展望

随着网络信息安全领域的发展，图像篡改检测技术也受到越来越多的重视，针对目前领域存在的问题，本文对其未来可能发展的方向进行了总结：一方面，基于在线社交网络的大背景，针对在社交网络中多次传输的篡改图像，提高算法在不同攻击下的鲁棒性是一个值得研究的发展方向。另一方面，鉴于 Transformer 方法中自注意力机制在图像篡改检测领域的优势，以及少数基于 Transformer 的方法已取得了不错的性能表现，因此将 Transformer 框架引入图像篡改检测领域具有广阔的应用前景。此外，由于创建大规模的基准数据集并非短时间能完成的工作，因此研究少样本学习、弱监督学习和迁移学习等方法在图像篡改检测领域的应用是一个可以深入研究的方向。最后，随着 AI(artificial intelligence) 技术的迅速发展，一些不法分子通过 AIGC(AI generated content) 生成更加逼真的篡改图像，如 AI 更换人脸等，因此针对 AIGC 生成的伪造图像进行方法技术研究将会是图像篡改领域在未来的重要方向之一。

4　结束语

本文对基于深度学习的图像篡改检测技术进行了总结，首先对目前的图像篡改检测领域的研究现状进行了介绍，其次在不同的角度对基于深度学习的篡改检测方法进行了分类总结，然后介绍了相关的数据集和评估指标并对不同方法的性能进行了直观对比，最后指出了目前方法的局限性并对未来的发展方向进行了展望。

参考文献：

[1] 田秀霞, 李华强, 张琴, 等. 基于双通道 R-FCN 的图像篡改检测模型[J]. 计算机学报, 2021, 44(2): 370−383.
TIAN Xiuxia, LI Huaqiang, ZHANG Qin, et al. Dual-channel R-FCN model for image forgery detection[J]. Chinese journal of computers, 2021, 44(2): 370−383.

[2] 胡林辉, 陈保营, 谭舜泉, 等. 基于 Convnext-Upernet 的图像篡改检测定位模型[J]. 计算机学报, 2023, 46(10): 2225−2239.
HU Linhui, CHEN Baoying, TAN Shunquan, et al. Convnext-upernet based deep-learning model for image forgery detection and localization[J]. Chinese journal of computers, 2023, 46(10): 2225−2239.

[3] 蔺琛皓, 沈超, 邓静怡, 等. 虚假数字人脸内容生成与检测技术[J]. 计算机学报, 2023, 46(3): 469−498.
LIN Chenhao, SHEN Chao, DENG Jingyi, et al. False digital facial content generation and detection technology[J]. Journal of computer science, 2023, 46(3): 469−498.

[4] 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 Haodong, ZHUANG Peiyu, LI Bin. A survey on deep learning based digital image tampering localization methods[J]. Journal of signal processing, 2021, 37(12): 2278−2301.

[5] 孙鹏, 郎宇博, 樊舒, 等. 图像拼接篡改的自动色温距离分类检验方法[J]. 自动化学报, 2018, 44(7): 1321−1332.
SUN Peng, LANG Yubo, FAN Shu, et al. Detection of image splicing manipulation by automated classification of color temperature distance[J]. Acta automatica sinica, 2018, 44(7): 1321−1332.

[6] 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 Xiaolong, YU Nenghai, ZHANG Xinpeng, et al. Overview of digital media forensics technology[J]. Journal of image and graphics, 2021, 26(6): 1216−1226.

[7] POPESCU A C, FARID H. Exposing digital forgeries in color filter array interpolated images[J]. IEEE transactions on signal processing, 2005, 53(10): 3948−3959.

[8] FARID H, LYU Siwei. Higher-order wavelet statistics and their application to digital forensics[C]//2003 Conference on Computer Vision and Pattern Recognition Workshop. Madison: IEEE, 2003: 94.

[9] FRIDRICH J, SOUKAL D, LUKAS J. Detection of copy-move forgery in digital images[C]//Proceedings of digital forensic research workshop. Cleveland: Elsevier, 2003: 652−662.

[10] 刘丽颖, 王金鑫, 曹少丽, 等. 检测小篡改区域的 U 型网络[J]. 中国图象图形学报, 2022, 27(1): 176−187.
LIU Liying, WANG Jinxin, CAO Shaoli, et al. U-Net for detecting small forgery region[J]. Chinese journal of image and graphics, 2022, 27(1): 176−187.

[11] 陈海鹏, 张世博, 吕颖达. 多尺度感知与边界引导的图像篡改检测方法[J/OL]. 吉林大学学报 (工学版). [2024−03−04]. https://doi.org/10.13229/j.cnki.jdxbgxb.20231027.
CHEN Haipeng, ZHANG Shibo, LYU Yingda. Multi-scale context-aware and boundary-guided image manipulation detection method[J/OL]. Journal of Jilin University (engineering edition). [2024−03−04]. https://doi.org/10.13229/j.cnki.jdxbgxb.20231027.

[12] 魏伟一, 赵毅凡, 陈帼. 基于深度特征提取和 DCT 变换的图像复制粘贴篡改检测[J]. 计算机工程与科学, 2023, 45(1): 163−170.
WEI Weiyi, ZHAO Yifan, CHEN Guo. Image copy-move forgery detection based on deep feature extraction and DCT transform[J]. Computer engineering and science, 2023, 45(1): 163−170.

[13] 朱新同, 唐云祁, 耿鹏志. 数字图像篡改检测技术综述[J]. 中国人民公安大学学报 (自然科学版), 2022, 28(4): 87−99.
ZHU Xintong, TANG Yunqi, GENG Pengzhi. Survey on digital image tampering detection technology[J]. Journal of People's Public Security University of China (natural science edition), 2022, 28(4): 87−99.

[14] 左鑫兰. 图像篡改检测技术研究综述[J]. 长江信息通信, 2022, 35(3): 74−76.
ZUO Xinlan. Research of image manipulation detection technology[J]. Changjiang information and communication, 2022, 35(3): 74−76.

[15] BI Xiuli, WEI Yang, XIAO Bin, et al. RRU-Net: the ringed residual U-Net for image splicing forgery detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach: IEEE, 2019: 30−39.

[16] VERDOLIVA L. Media forensics and DeepFakes: an overview[J]. IEEE journal of selected topics in signal processing, 2020, 14(5): 910−932.

[17] MEHRJARDI F Z, LATIF A M, ZARCHI M S, et al. A survey on deep learning-based image forgery detection[J]. Pattern recognition, 2023, 144: 109778.

[18] PHAM N T, PARK C S. Toward deep-learning-based methods in image forgery detection: a survey[J]. IEEE access, 2023, 11: 11224−11237.

[19] ZANARDELLI M, GUERRINI F, LEONARDI R, et al. Image forgery detection: a survey of recent deep-learning approaches[J]. Multimedia tools and applications, 2023, 82(12): 17521−17566.

[20] 杨衍宇, 魏为民, 张运琴. 一种改进的双流 Faster R-CNN 图像篡改识别模型[J]. 计算机应用与软件, 2023, 40(12): 189−194.
YANG Yanyu, WEI Weimin, ZHANG Yunqin. Image forgery recognition model based on improved dual-stream faster R-CNN[J]. Computer applications and software, 2023, 40(12): 189−194.

[21] 付顺旺, 陈茜, 李智, 等. 用于篡改图像检测和定位的双通道渐进式特征过滤网络[J]. 计算机应用, 2024, 44(4): 1303−1309.
FU Shunwang, CHEN Qian, LI Zhi, et al. Two-channel progressive feature filtering network for tampered image detection and localization[J]. Journal of computer applications, 2024, 44(4): 1303−1309.

[22] ZHOU Peng, HAN Xintong, MORARIU V I, et al. Learning rich features for image manipulation detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 1053−1061.

[23] ZHOU Peng, HAN Xintong, MORARIU V I, et al. Two-stream neural networks for tampered face detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017: 1831−1839.

[24] REN Shaoqing, HE Kaiming, 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, 2016, 39(6): 1137−1149.

[25] RAO Yuan, NI Jiangqun, ZHANG Weizhe, et al. Towards JPEG-resistant image forgery detection and localization via self-supervised domain adaptation[J]. IEEE transactions on pattern analysis and machine intelligence, 2022: 1−12.

[26] CHEN Xinru, DONG Chengbo, JI Jiaqi, et al. Image manipulation detection by multi-view multi-scale supervision[C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 14165−14173.

[27] LI Dong, ZHU Jiaying, WANG Menglu, et al. Edge-aware regional message passing controller for image forgery localization[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 8222−8232.

[28] LIU Yaqi, LYU Binbin, JIN Xin, et al. TBFormer: two-branch Transformer for image forgery localization[J]. IEEE signal processing letters, 2023, 30: 623−627.

[29] BAYAR B, STAMM M C. Constrained convolutional neural networks: a new approach towards general purpose image manipulation detection[J]. IEEE transactions on information forensics and security, 2018, 13(11): 2691−2706.

[30] FU Jun, LIU Jing, TIAN Haijie, et al. Dual attention network for scene segmentation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 3146−3154.

[31] WU Yue, ABDALMAGEED W, NATARAJAN P. ManTra-net: manipulation tracing network for detection and localization of image forgeries with anomalous features [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 9543−9552.

[32] LIN Xun, WANG Shuai, DENG Jiahao, et al. Image manipulation detection by multiple tampering traces and edge artifact enhancement[J]. Pattern recognition, 2023, 133: 109026.

[33] DAS S, ISLAM M S, AMIN M R. GCA-net: utilizing gated context attention for improving image forgery localization and detection[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans: IEEE, 2022: 81−90.

[34] WARIF N B A, IDRIS M Y I, WAHAB A W A, et al. An evaluation of error level analysis in image forensics[C]// 2015 5th IEEE International Conference on System Engineering and Technology. Shah Alam: IEEE, 2015: 23−28.

[35] GUILLARO F, COZZOLINO D, SUD A, et al. TruFor: leveraging all-round clues for trustworthy image forgery detection and localization[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 20606−20615.

[36] COZZOLINO D, VERDOLIVA L. Noiseprint: a CNN-based camera model fingerprint[J]. IEEE transactions on information forensics and security, 2019, 15: 144−159.

[37] WANG Junke, WU Zuxuan, CHEN Jingjing, et al. ObjectFormer for image manipulation detection and localization[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 2354−2363.

[38] GUO Xiao, LIU Xiaohong, REN Zhiyuan, et al. Hierarchical fine-grained image forgery detection and localization[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 3155−3165.

[39] KWON M J, NAM S H, YU I J, et al. Learning JPEG compression artifacts for image manipulation detection and localization[J]. International journal of computer vision, 2022, 130(8): 1875−1895.

[40] ZHANG Zhenfei, LI Mingyang, CHANG M C. A new benchmark and model for challenging image manipulation detection[J]. Proceedings of the AAAI conference on artificial intelligence, 2024, 38(7): 7405−7413.

[41] GU A R, NAM J H, LEE S C. FBI-net: frequency-based image forgery localization via multitask learning with self-attention[J]. IEEE access, 2022, 10: 62751−62762.

[42] LIU Yaqi, ZHU Xiaobin, ZHAO Xianfeng, et al. Adversarial learning for constrained image splicing detection and localization based on atrous convolution[J]. IEEE transactions on information forensics and security, 2019, 14(10): 2551−2566.

[43] NILOY F F, KUMAR BHAUMIK K, WOO S S. CFL-net: image forgery localization using contrastive learning [C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2023: 4631−4640.

[44] ZHONG Junliu, PUN C M. An end-to-end dense-InceptionNet for image copy-move forgery detection[J]. IEEE transactions on information forensics and security, 2019, 15: 2134−2146.

[45] HAO Jing, ZHANG Zhixin, YANG Shicai, et al. TransForensics: image forgery localization with dense self-attention[C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 15035−15044.

[46] LIU Ze, LIN Yutong, CAO Yue, et al. Swin Transformer: hierarchical vision Transformer using shifted windows[C]// 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 9992−10002.

[47] TOUVRON H, CORD M, SABLAYROLLES A, et al. Going deeper with image Transformers[C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 32−42.

[48] ZHOU Zongwei, RAHMAN SIDDIQUEE M M, TAJBAKHSH N, et al. UNet++: a nested U-Net architecture for medical image segmentation[M]//Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Cham: Springer International Publishing, 2018: 3−11.

[49] LIU Xiaohong, LIU Yaojie, CHEN Jun, 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.

[50] WANG Jingdong, SUN Ke, CHENG Tianheng, et al. Deep high-resolution representation learning for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 43(10): 3349−3364.

[51] SALLOUM R, REN Yuzhuo, 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.

[52] SUN Zhihao, JIANG Haoran, WANG Danding, et al. SAFL-net: semantic-agnostic feature learning network with auxiliary plugins for image manipulation detection [C]//2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 22367−22376.

[53] ZHOU Peng, CHEN B C, HAN Xintong, et al. Generate, segment, and refine: towards generic manipulation segmentation[J]. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(7): 13058−13065.

[54] Ma Xiaochen, Du Bo, Liu Xianggen, et al. IML-ViT: Image manipulation localization by vision Transformer [EB/OL]. (2023−07−27)[2024−03−04]. https://arxiv.org/abs/2307.14863v1.

[55] ZENG Yuyuan, ZHAO Bowen, QIU Shanzhao, et al. Toward effective image manipulation detection with proposal contrastive learning[J]. IEEE transactions on circuits and systems for video technology, 2023, 33(9): 4703−4714.

[56] HAO Qixian, REN Ruyong, WANG Kai, et al. EC-Net: General image tampering localization network based on edge distribution guidance and contrastive learning[J]. Knowledge-based systems, 2024, 293: 111656.

[57] WU Haiwei, CHEN Yiming, ZHOU Jiantao. Rethinking image forgery detection via contrastive learning and unsupervised clustering[EB/OL]. (2023−08−18) [2024−03−04]. https://arxiv.org/abs/2308.09307v1.

[58] ZHOU Jizhe, MA Xiaochen, DU Xia, et al. Pre-training-free image manipulation localization through non-mutually exclusive contrastive learning[C]//2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 22289−22299.

[59] BARNI M, PHAN Q T, TONDI B. Copy move source-target disambiguation through multi-branch CNNs[J]. IEEE transactions on information forensics and security, 2020, 16: 1825−1840.

[60] ZHANG Ying, GOH J, WIN Lei Lei, et al. Image region forgery detection: a deep learning approach[C]//Proceedings of the Singapore Cyber-Security Conference 2016. Singapore: IOS Press, 2016: 1−11.

[61] ISLAM A, LONG Chengjiang, BASHARAT A, et al. DOA-GAN: dual-order attentive generative adversarial network for image copy-move forgery detection and localization[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 4676−4685.

[62] 石泽男, 陈海鹏, 张冬, 等. 预训练驱动的多模态边界感知视觉 Transformer[J]. 软件学报, 2023, 34(5): 2051−2067.
SHI Zenan, CHEN Haipeng, ZHANG Dong, et al. Pre-training-driven multimodal boundary-aware vision Transformer[J]. Journal of software, 2023, 34(5): 2051−2067.

[63] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. (2020−10−22) [2024−03−04]. https://arxiv.org/abs/2010.11929.

[64] LI Yanghao, XIE Saining, CHEN Xinlei, et al. Benchmarking detection transfer learning with vision Transformers[EB/OL]. (2021−11−22)[2024−03−04]. https://arxiv.org/abs/2111.11429v1.

[65] LI Yanghao, WU Chaoyuan, FAN Haoqi, et al. MViTv2: improved multiscale vision Transformers for classification and detection[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 4794−4804.

[66] LI Yuxi, CHENG Fuyuan, YU Wangbo, et al. AdaIFL: adaptive image forgery localization via a dynamic and importance-aware Transformer network[C]//Computer Vision–ECCV 2024. Cham: Springer Nature Switzerland, 2024: 477−493.

[67] LI Shuaibo, MA Wei, GUO Jianwei, et al. UnionFormer: unified-learning Transformer with multi-view representation for image manipulation detection and localization [C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 12523−12533.

[68] ROZANTSEV A, SALZMANN M, FUA P. Beyond sharing weights for deep domain adaptation[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 41(4): 801−814.

[69] ZHUANG Peiyu, LI Haodong, YANG Rui, et al. ReLoc: a restoration-assisted framework for robust image tampering localization[J]. IEEE transactions on information forensics and security, 2023, 18: 5243−5257.

[70] WU Haiwei, ZHOU Jiantao, TIAN Jinyu, et al. Robust image forgery detection over online social network shared images[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 13430−13439.

[71] HSU Y F, CHANG S F. Detecting image splicing using geometry invariants and camera characteristics consistency[C]//2006 IEEE International Conference on Multimedia and Expo. Toronto: IEEE, 2006: 549−552.

[72] AMERINI I, BALLAN L, CALDELLI R, et al. A SIFT-based forensic method for copy–move attack detection and transformation recovery[J]. IEEE transactions on information forensics and security, 2011, 6(3): 1099−1110.

[73] DONG Jing, WANG Wei, TAN Tieniu. CASIA image tampering detection evaluation database[C]//2013 IEEE

China Summit and International Conference on Signal and Information Processing. Beijing: IEEE, 2013: 422−426.

[74] DE CARVALHO T J, 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.

[75] TRALIC D, ZUPANCIC I, GRGIC S, et al. CoMoFoD: New database for copy-move forgery detection[C]//Proceedings ELMAR-2013. Zadar: IEEE, 2013: 49−54.

[76] SILVA E, CARVALHO T, FERREIRA A, et al. Going deeper into copy-move forgery detection: Exploring image telltales via multi-scale analysis and voting processes [J]. Journal of visual communication and image representation, 2015, 29: 16−32.

[77] COZZOLINO D, POGGI G, VERDOLIVA L. Efficient dense-field copy–move forgery detection[J]. IEEE transactions on information forensics and security, 2015, 10(11): 2284−2297.

[78] WEN Bihan, ZHU Ye, SUBRAMANIAN R, et al. COVERAGE: a novel database for copy-move forgery detection[C]//2016 IEEE International Conference on Image Processing. Phoenix: IEEE, 2016: 161−165.

[79] ZAMPOGLOU M, PAPADOPOULOS S, KOMPATSIARIS Y. Detecting image splicing in the wild (WEB)[C]//2015 IEEE International Conference on Multimedia & Expo Workshops. Turin: IEEE, 2015: 1−6.

[80] KORUS P, HUANG Jiwu. Evaluation of random field models in multi-modal unsupervised tampering localization[C]//2016 IEEE International Workshop on Information Forensics and Security. Abu Dhabi: IEEE, 2016: 1−6.

[81] GUAN Haiying, KOZAK M, ROBERTSON E, et al. MFC datasets: large-scale benchmark datasets for media forensic challenge evaluation[C]//2019 IEEE Winter Applications of Computer Vision Workshops. Waikoloa Village: IEEE, 2019: 63−72.

[82] HELLER S, ROSSETTO L, SCHULDT H. The PS-Battles dataset—an image collection for image manipulation detection[EB/OL]. (2018−04−13)[2024−03−04]. https://arxiv.org/abs/1804.04866.

[83] MAHFOUDI G, TAJINI B, RETRAINT F, et al. DEFACTO: image and face manipulation dataset[C]//2019 27th European Signal Processing Conference. A Coruna: IEEE, 2019: 1−5.

[84] NOVOZAMSKY A, MAHDIAN B, SAIC S. IMD2020: a large-scale annotated dataset tailored for detecting manipulated images[C]//2020 IEEE Winter Applications of Computer Vision Workshops. Snowmass Village: IEEE, 2020: 71−80.

[85] WU Yue, ABD-ALMAGEED W, NATARAJAN P. BusterNet: detecting copy-move image forgery with source/target localization[C]//Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018: 170−186.

[86] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 740−755.

[87] XIAO Jianxiong, HAYS J, EHINGER K A, et al. SUN database: large-scale scene recognition from abbey to zoo[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco: IEEE, 2010: 3485−3492.

[88] GLOE T, BÖHME R, GLOE T, et al. The ‘Dresden Image Database’ for benchmarking digital image forensics [C]//Proceedings of the 2010 ACM Symposium on Applied Computing. Sierre: ACM, 2010: 1584−1590.

[89] ZHUANG Peiyu, LI Haodong, TAN Shunquan, et al. Image tampering localization using a dense fully convolutional network[J]. IEEE transactions on information forensics and security, 2021, 16: 2986−2999.

作者简介：

张汝波，教授，辽宁省教学名师，主要研究方向为智能机器人决策与控制技术。主持国家重点基础研究发展计划项目、国家高技术研究发展计划项目、国家自然科学基金项目等 20 余项。获得国家科学技术进步二等奖 1 项、省部级科学技术奖 6 项，获发明专利授权 10 余项。发表学术论文 200 余篇。E-mail：zhangrubo@dlnu.edu.cn。

蔺庆龙，硕士研究生，主要研究方向为图像篡改检测。E-mail：linqinglong1999@163.com。

张天一，助理教授，博士，主要研究方向为计算机视觉、机器学习以及图像视频内容安全分析，具体研究内容包括图像分割、目标检测、视频动作识别、深度学习、弱监督学习。主持国家自然科学基金项目 1 项、企事业单位委托项目 1 项。发表学术论文 10 余篇。E-mail：zhang_tianyi@buaa.edu.cn。