

A Survey of Image Forgery Detection and Localization

WANG Wei*, LI Ang, NI Rongrong, WEI Yunchao, ZHAO Yao

(School of Computer Science and Technology, Beijing Jiaotong University, Beijing 100044, China)

Abstract: Image forgery detection and localization (IFDL) determines whether an image has been manipulated or synthesized in its entirety and identifies locally manipulated regions. Generative editing and whole-image synthesis change the available forensic traces and broaden the scope of detection research. This survey reviews representative literature from 2005 to 2026. Conventional local manipulation, generative local editing and fully synthetic images are distinguished by how their pixels are produced. Their traces, datasets and evaluation metrics are summarized. Representative methods are then discussed along three research directions: conventional forensic cues, generative forgery analysis and multimodal large language models. The review compares their mechanisms, applicable conditions, evaluation settings and performance under transmission degradation. The analysis shows that image-level labels and pixel-level masks support different evaluation tasks, while the reported composition of genuine and forged images is inconsistent across publications for some datasets. Training data, supervision, input resolution and threshold selection affect performance comparisons; some differences in evaluation settings also change method rankings. Under generative editing, classical cues depend on background preservation, region blending and subsequent processing. Generator artifacts, diffusion-inversion residuals and pretrained features provide complementary evidence, but generalization remains sensitive to differences between generators and data distributions. Multimodal large language models introduce natural-language explanations, yet methods differ in their support for localization masks and source attribution. Explanation readability and faithfulness to visual evidence require separate assessment. The effects of compression, resizing and noise are related to the cues used; controlled experiments are still needed to distinguish the contributions of training data and network architecture. This survey provides a reference for selecting forensic evidence, designing comparable evaluations and choosing methods for practical deployment.

Highlights:

1. Forensic cues are compared across conventional manipulation, generative editing and image synthesis.
2. Detection, localization, attribution and explanation are distinguished in evaluating methods.
3. Evaluation settings and transmission degradation are examined as factors affecting performance.

Key words: image forgery detection; forgery localization; diffusion models; multimodal large language models; forensic cues; evaluation protocols

Foundation items: National Natural Science Foundation of China (No.62372033); Natural Science Foundation of Beijing, China (No.L252025); Project of ‘111 Center’ (No.B25028).

Received: 2026-08-15; **Revised:** 2026-09-11

***Corresponding author, E-mail:** wei.wang@bjtu.edu.cn.

图像伪造检测与定位技术综述

王伟, 李昂, 倪蓉蓉, 魏云超, 赵耀

(北京交通大学计算机科学与技术学院, 北京 100044)

摘要: 图像伪造检测与定位(Image forgery detection and localization, IFDL)旨在判别图像是否经过篡改或整幅合成,并定位局部篡改区域。生成式编辑与整幅合成改变了可用取证痕迹,也扩展了检测方法的研究范围。本文梳理了2005年至2026年的代表性文献,按像素产生机制区分传统局部篡改、生成式局部编辑与全合成图像,整理其痕迹、数据集及评价指标;围绕传统取证线索、生成式伪造与多模态大模型3条研究路线,介绍代表方法的机制与适用条件,比较评测设置及传输退化下的性能。分析表明,图像级标签与像素级掩膜支持不同评测任务,部分数据集的真伪构成在各文献中记载不一;训练数据、监督方式、输入分辨率与阈值选取影响性能比较,部分设置差异还会改变方法排名。经典线索在生成式编辑中的有效性取决于背景保留、区域融合及后处理,生成器伪影、扩散反演残差与预训练特征提供了补充依据,但泛化能力仍受生成器和数据分布差异影响。多模态大模型引入自然语言解释,各方法对掩膜与来源归因的支持程度不同,解释的可读性与证据忠实性也需分别评价。压缩、缩放与噪声的影响与所用线索有关,训练数据和网络架构的作用尚需受控于实验区分。本文为取证依据选择、可比评测与实际部署提供参考。

关键词: 图像伪造检测;伪造区域定位;扩散模型;多模态大模型;取证线索;评测协议

中图分类号: TP391.41

文献标志码: A

引用格式: 王伟,李昂,倪蓉蓉,等.图像伪造检测与定位技术综述[J].数据采集与处理,2026,41(5):1268-1305.
WANG Wei, LI Ang, NI Rongrong, et al. A survey of image forgery detection and localization[J]. Journal of Data Acquisition and Processing, 2026, 41(5):1268-1305.

引言

扩散模型与指令式编辑工具降低了图像伪造的制作门槛。在司法举证、新闻取证、保险理赔与社交平台内容治理中,图像核验既要判断是否存在伪造,也要确定改动位置。人工智能生成内容(AI-generated content, AIGC)的标识要求进一步增加了图像伪造检测与定位(Image forgery detection and localization, IFDL)的应用需求。

早期研究从成像与编辑过程的统计规律中提取判别依据。Popescu等^[1]利用重采样引入的周期性相关检测缩放或旋转痕迹。Fridrich等^[2]提出的富模型通过高通残差的高阶共生统计描述图像,减少对特定操作痕迹的依赖。Cozzolino等^[3]提出的Splicebuster采用无监督的期望最大化算法,根据局部共生特征分割单幅图像中的统计异常区域。这些方法的适用范围取决于相应痕迹的形成与保留条件。

卷积网络将取证特征的获取扩展到端到端学习。Zhou等^[4]提出的RGB-N在目标检测框架中融合红绿蓝(Red-green-blue, RGB)图像流与富模型噪声流;Chen等^[5]提出的MVSS-Net联合学习边缘与噪

声特征,并引入真实图像负样本降低误报;Kwon等^[6]提出的CAT-Net则利用离散余弦变换系数与量化表学习压缩历史差异。多类特征的融合扩大了可处理的篡改范围,但跨库表现仍受训练数据分布影响。

相关研究还扩展了取证输出。Guillaro等^[7]提出的TruFor以置信图估计局部定位结果的可靠性;Guo等^[8]提出的HiFi-Net通过分层分类实现真伪判定、掩膜预测与细粒度归因;Xu等^[9]提出的FakeShield面向传统及生成式篡改,在输出判定与掩膜的同时生成视觉证据的文字解释。

尽管相关方法的判别依据和输出形式不断演进,该领域仍面临两方面困难。其一是跨域比较不足。ForensicHub^[10]将Deepfake、图像篡改、AIGC与文档取证纳入统一评测,指出这些方向的数据与评价标准存在差异,专门领域内的优势未必能够迁移。其二是经典线索的适用条件发生变化。依赖区域间不一致的方法通常要求篡改区域与原背景区域衔接处保留可分的成像差异或处理痕迹,生成式编辑与后处理可能削弱这些差异。影响程度取决于是否保留原始背景、如何融合编辑区以及后续压缩强度,不能由生成模型的使用直接推定线索已经消失。

已有综述从不同任务范围整理了相关进展。面向媒体取证,Verdoliva^[11]系统介绍了视觉媒体完整性验证的主要问题与技术。面向图像篡改检测与定位,Kadha等^[12]按篡改类型归纳手工特征与卷积网络方法,张汝波等^[13]则根据多流融合、多尺度处理等共性技术梳理篡改定位方法。面向生成内容检测,李美玲等^[14]以图像级二分类为对象介绍人工智能生成图像检测技术,多模态大模型相关综述^[15]进一步覆盖多种模态的生成内容,像素级定位是其中一项任务。彭春蕾等^[16]关注大模型时代的深度伪造检测,刘晓龙等^[17]兼顾被动检测与主动防御,两者对像素级篡改定位的系统比较相对有限。可解释性相关综述^[18]则从定位、归因与解释3个维度,结合4类特征策略组织研究进展。这些综述的任务范围与分类方式各有侧重,对生成式局部编辑及其对经典取证线索适用性的影响仍有进一步梳理的空间。本文据此以取证线索为主线,介绍各类方法的机制、适用条件及失效原因,并比较其图像级检测与像素级定位能力,所覆盖的方法自2005年的重采样检测起,延至2026年。

本文覆盖传统局部篡改、生成式局部编辑与全合成图像,以取证线索为主线梳理数据与方法,技术框架如图1所示。全文先界定任务与伪造源,介绍痕迹、数据集及评价指标,再比较传统取证、生成式伪造分析与多模态方法,最后讨论评测设置、传输链路鲁棒性及研究挑战。

1 图像伪造检测与定位的任务与痕迹谱系

图像级检测、像素级定位、来源归因与自然语言解释对应不同的输出任务,其适用条件与伪造区域的像素来源有关。本节据此界定任务范围,介绍3类伪造源及其痕迹,并说明任务之间的关系与常见处理流程。

1.1 任务定义与术语辨析

典型的被动图像取证以来源与处理历史未知的图像为输入,在没有参考图像或编辑记录的条件判断是否存在伪造,并分析其位置与形成方式。为比较不同工作的任务范围,本文将已有输出整理为4个维度,各方法可实现其中的一项或多项:(1)图像级真伪判定,给出真伪标签或伪造置信分数,可用于平台侧初筛;(2)像素级篡改掩膜,以同尺寸概率图标出改动区域,为司法举证与内容溯源提供位置信息;(3)篡改类型与生成器归因,HiFi-Net^[8]采用分层分类树逐级预测来源;(4)自然语言解释,FakeShield^[9]将视觉证据的文字说明纳入任务定义。

两项统一基准分别采用图像篡改检测与定位(Image manipulation detection and localization, IMDL)

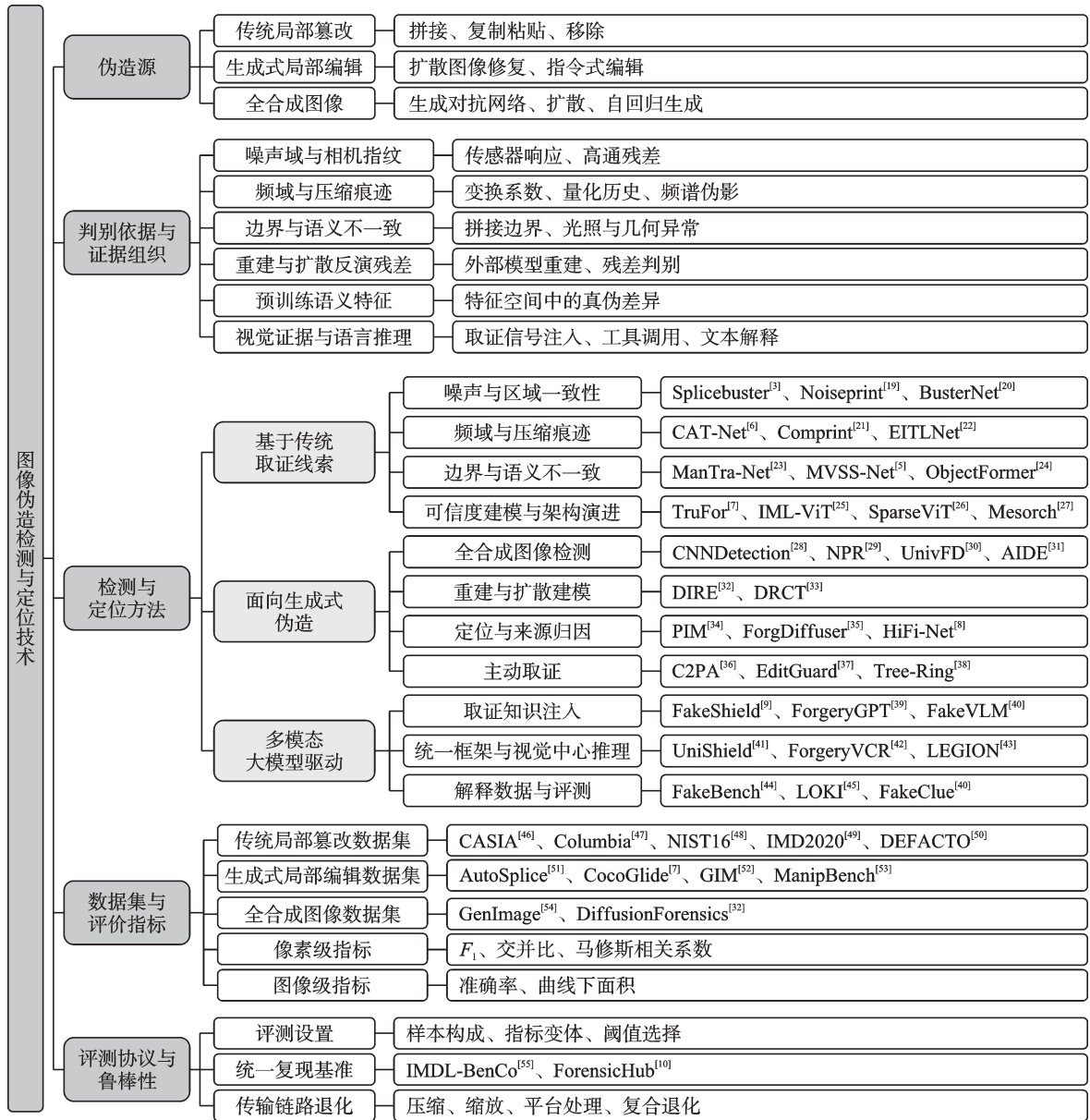


图1 图像伪造检测与定位技术框架

Fig.1 Technical framework of image forgery detection and localization

与假图检测与定位(Fake image detection and localization, FIDL)命名,其研究范围有所不同。IMDL-BenCo^[55]所用的IMDL偏重传统局部篡改;ForensicHub所用的FIDL同时涵盖Deepfake、图像篡改、AIGC与文档取证,该基准指出4个方向的数据与评价标准存在差异,限制了跨域比较^[10]。本文沿用HiFi-Net中IFDL的名称,讨论传统篡改与整幅生成图像,不系统展开人脸Deepfake检测^[11],仅在跨域评测及可解释方法中介绍相关代表工作。因此,本文范围比IMDL-BenCo更广、比ForensicHub更窄;这一比较针对具体研究范围,3个缩写本身并不存在普遍固定的包含关系。

任务名称包含“检测”,并不意味着每个方法都输出图像级真伪判定。IMDL-BenCo统一复现的8

个代表性方法中,参与图像级评测的只有3个^[55],分别为MVSS-Net^[5]、PSCC-Net^[56]与TruFor^[7]。其余5个方法在该基准的实现中未配置专门的检测头,只参加像素级定位评测。不同方法对这4个维度的支持程度有所区别,评价指标也对应不同的输出任务。部分方法只以边界框标出可疑区域,其定位粒度与像素级掩膜不同。仅输出掩膜的方法与支持3项以上输出的方法,可以比较其共同支持的任务,单项排名则不能代表整体取证能力。

1.2 伪造源的分类

相同的内容修改可以通过不同操作实现,例如人脸替换既可采用像素复制,也可采用扩散重绘,两者留下的痕迹不同。因此,本文按篡改区像素的产生机制区分3类伪造源;这一分类描述图像的产生方式,与检测方法的技术路线并非一一对应。

(1)传统局部篡改,包括拼接、复制粘贴与采用传统修补算法的移除,主要通过像素复制、重组或邻域估计改变内容,并辅以模糊、调色等后处理。拼接可能将不同来源的传感器噪声与压缩历史带入同一幅图,Splicebuster^[3]一类无监督检测器利用区域统计差异进行定位;同图复制则未必具有相机来源差异,还需分析重复内容、边界或重采样痕迹。代表数据集为CASIA^[46]、COVERAGE^[57]与NIST16^[48]等人工标注的小规模数据集。这一类以已有拍摄内容及传统修补为基础,不依靠生成模型合成新的局部内容。

(2)生成式局部编辑,其篡改区的像素由扩散模型等生成模型合成:扩散图像修复按给定掩膜重绘局部内容,指令式编辑则按一句自然语言指令改写指定对象。部分流程会使背景也经过编码与解码,另一些流程则将生成区域与原始背景合成;区域统计与边界痕迹是否保留,取决于具体的编辑及融合方式。GIM^[52]以全自动流水线批量制作此类图像,ManipBench则覆盖25个编辑模型与12类编辑操作^[53]。这类编辑的痕迹受到生成模型与后处理影响;与全合成图像相比,它以局部内容为编辑目标,篡改掩膜用于标注目标区域,背景是否逐像素保留取决于具体实现。

(3)全合成图像,整幅画面由生成对抗网络、扩散模型或自回归模型生成。由于未经历真实相机成像过程,这类图像缺少相应的传感器与色彩滤波阵列痕迹,图内也没有原始拍摄区域可供参照。检测可利用上采样周期栅格等生成器伪影,也可分析图像内容及统计分布异常。标准评测集GenImage^[54]以百万级规模覆盖8个生成器。与局部编辑相比,全合成图像的生成范围覆盖整幅图像,真值掩膜因而为全图常量,不能用于区分图内的编辑区域,评测重点相应转向图像级真伪判定。

3类伪造源的对照如表1所示。局部篡改可能保留原始背景,支持区域间比较;全合成图像缺少这一参照,更多依赖生成器伪影及内容异常。表中工具与数据集取自各自原文,痕迹特点按成像与生成机制归纳。

表1 3类伪造源对照

Table 1 Comparison of three types of forgery sources

类型	生成方式	典型工具与数据集	痕迹特点
传统局部篡改	像素复制、传统修补、后处理	CASIA ^[46] 、COVERAGE ^[57] 、NIST16 ^[48]	区域统计或边界异常
生成式局部编辑	按掩膜重绘、按文本改写	Stable Diffusion ^[58] 、GLIDE ^[59] 、DDNM ^[60] 、CocoGlide ^[7] 、GIM ^[52] 、ManipBench ^[53]	区域差异依赖融合方式
全合成图像	整幅生成,无拍摄内容	ProGAN ^[61] 、Stable Diffusion ^[58] 、GenImage ^[54]	无传感器痕迹、图内无参照区

1.3 痕迹谱系与经典线索的失效机理

数字图像的采集与编辑可能涉及光电转换、色彩滤波阵列插值、有损编码、几何变换与区域合成,各环节会留下不同的统计痕迹。按痕迹的形成阶段,可以将取证线索整理为谱系,并分析其适用条件。传感器成像阶段的像素响应非均匀性(Photo-response non-uniformity, PRNU)表征设备的传感器响应差异,但其估计需要同一台设备的多张参考图像进行对齐平均,网络图像通常难以满足这一条件^[62]。几何变换中的插值则使邻近像素产生周期性线性相关,在二阶导数频谱上表现为周期峰。Popescu等^[1]指出,在所测条件下,当联合图像专家组(Joint photographic experts group, JPEG)压缩的质量因子低于约95时,这些峰可能被量化噪声淹没。

量化与双压缩痕迹反映了图像的编码历史。一幅JPEG图像被解码、编辑、再存盘之后,若背景与粘贴区的初始压缩历史不同,未改动区域可能经历两次量化而篡改区只经历一次,块级分析据此可以给出每个8×8块的压缩次数后验概率^[63]。CAT-Net^[6]则将量化后的离散余弦变换(Discrete cosine transform, DCT)系数编码为二值系数体,在保留空间位置的基础上结合量化表学习压缩差异。合成与生成阶段也会产生痕迹:拼接边界的高频跃变可由MVSS-Net的边缘监督分支学习,部分生成器的上采样会留下周期栅格,Synthbuster^[64]先抑制图像内容再检测残差频谱的周期峰,对多数训练时未见过的扩散生成器可达到0.9以上的曲线下面积。除图像形成时留下的痕迹外,DIRE^[32]在检测时对待检图像进行扩散反演与重建,以原图和重建图的残差构成判别图。反演并非待检图像形成时必经的工序,其判别力取决于残差能否区分目标分布中的真实与伪造图像。图2将痕迹形成阶段与检测时构造的参照残差分别列示。

各类线索的适用性取决于其形成与保留条件。以区域间比较为依据的定位器,需要背景与编辑区仍存在可分的统计差异;若背景也被明显改写,这一参照可能减弱。Noiseprint^[19]所刻画的相机模型残

图像形成与处理痕迹					检测时参照
					扩散反演
传统局部篡改	传感器成像	插值重采样	JPEG量化与双压缩	合成边界	生成器解码
	条件可用	条件可用	条件可用	条件可用	不适用
	PRNU Noiseprint ^[19]	重采样 周期峰	块级双压缩 CAT-Net ^[6]	MVSS-Net ^[5] 边缘监督	生成器解码 未参与
	条件受限	条件受限	条件受限	条件受限	条件受限
生成式局部编辑	PRNU 局部被覆盖	重采样峰 随重绘变化	块级双压缩 编辑区失配	边界差异 随融合变化	生成谱峰 依赖具体模型
	条件受限	条件受限	条件受限	条件受限	条件受限
全合成图像	不适用	不适用	不适用	不适用	条件可用
	无原始 传感器痕迹	无原始 相机插值痕迹	无区域间 压缩史参照	全图合成 无拼接边界	Synthbuster ^[64] 谱峰
					条件可用
					DIRE ^[32] 反演残差

条件可用 条件受限 不适用

图2 伪造痕迹谱系(条件随编辑与后处理变化)

Fig.2 Forensic trace taxonomy under different editing and post-processing conditions

差可以用于区域一致性分析,但不能仅凭编辑流程包含解码器就认定其失去全部判别力。类似地,再压缩可能削弱量化历史差异,区域融合可能平滑边界,但影响并非对所有图像与方法相同。性能下降既可能源于输入线索减弱,也可能源于模型未能利用残余信息;背景保留方式、融合操作和后处理设置共同构成两种解释的验证条件。

1.4 任务能力之间的相互制约

各项输出的评价结果受到任务定义与伪造范围的共同影响。当整幅图像均由模型生成时,真值掩膜为全1矩阵,将所有像素判为伪造即可获得满分,交并比与像素级 F_1 分数难以继续区分定位能力。局部篡改即使只涉及少量像素,图像级标签也为伪造;但检测难度不只由篡改范围决定,图像级评测结果还受到真图负样本与阈值选取的影响。

在同一批模型权重上,检测与定位的排名也可能不同。IMDL-BenCo^[55]在以 MVSS-Net 原始设置为基准的训练协议 Protocol-MVSS 下,对 3 个同时提供检测输出的方法进行比较:TruFor 的像素级 F_1 为 0.530,居三者之首,图像级 F_1 为 0.598,居三者之末;MVSS-Net 的像素级均值排名第 2,图像级结果则以 0.631 居首。像素级均值涵盖 5 个测试集,图像级评测排除了 CASIAv1,只计算其余四库,因此这些排名还受到测试集构成的影响,不能单独归因于检测与定位的任务差异。

MVSS-Net 原文^[5]中,像素级排名第 2 的方法得分为 0.324,但同批图像级评测的特异度不超过 0.011、 F_1 分数不超过 0.022。该原文的图像级 F_1 定义为灵敏度与特异度的调和平均,与精确率和召回率定义的 F_1 不同。仅在伪造图像上统计像素级 F_1 时,完整真实图像上的假阳性不计入结果。因此,定位准确性与真伪判定能力需要分别评价,比较时还须核对指标定义与样本范围。

伪造源类型与任务能力的对应关系如图 3 所示。横轴为 3 类伪造源,格中状态表示任务的适用条件。即使仅有局部篡改,图像级标签也为伪造;全合成图像的所有像素均标为伪造,真值掩膜为全 1,不能据此评价局部篡改定位能力。

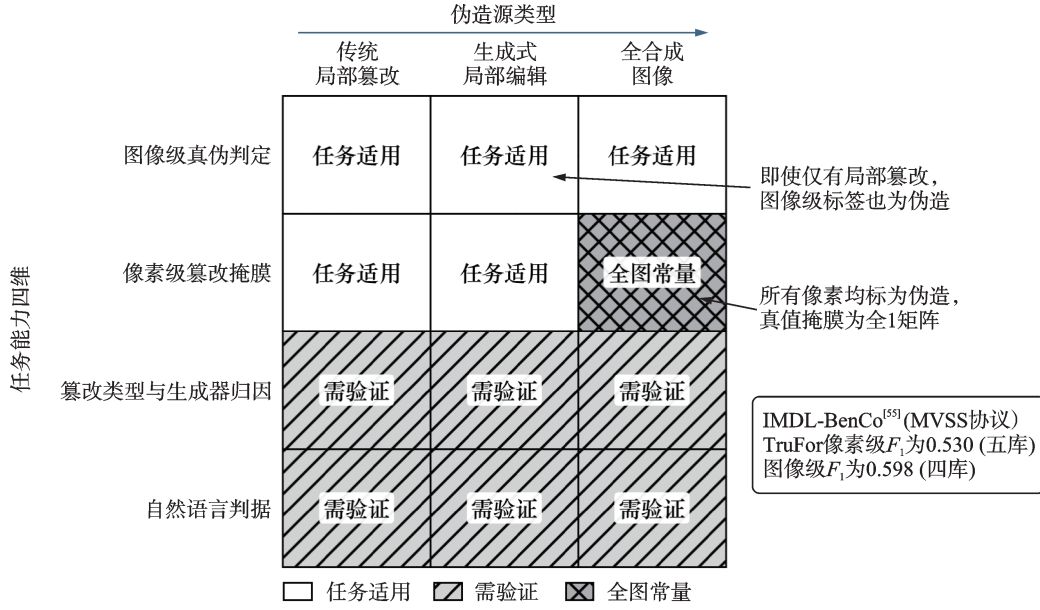
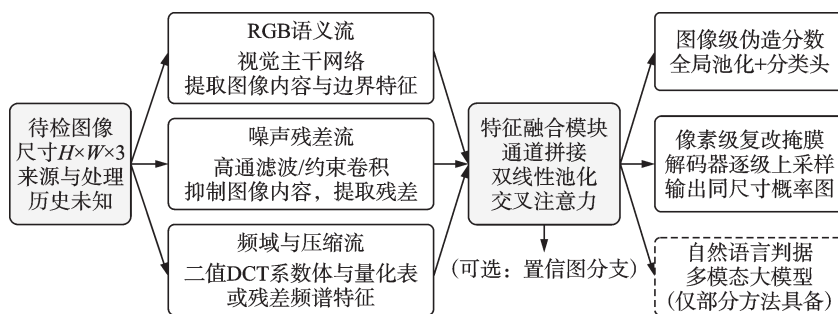


图 3 伪造源类型与任务能力的二维网络

Fig.3 Two-dimensional network of forgery sources and task capabilities

1.5 通用检测与定位流程

多种 IFDL 方法可归纳为特征提取、融合与预测 3 个环节,通用流程如图 4 所示。其中包括 3 条可选的特征提取支路:RGB 语义流从像素中提取内容与边界信息,噪声残差流抑制图像内容并提取与成像处理有关的残差,频域与压缩流则利用 DCT 系数及量化信息描述编码历史,或从残差频谱提取生成伪影。各支路特征在尺度对齐后,通过通道拼接或交叉注意力融合,再由解码器逐级上采样,输出与输入同尺寸的像素级篡改概率图。融合特征也可经全局池化与分类头生成图像级伪造分数,部分多模态方法进一步结合语言模型生成自然语言解释。图中模块可按方法选用,并非所有方法都采用完整流程。



注: H 、 W 分别为图像高度与宽度, RGB 为红绿蓝图像通道; 各特征支路按具体方法选用。

图 4 图像伪造检测与定位的通用流程

Fig.4 General pipeline for image forgery detection and localization

2 数据集与评价指标

数据集决定训练与测试中的伪造类型、图像分辨率及后处理条件,评价指标衡量特定输出任务的表现。本节按 3 类伪造源介绍常用数据集,并说明像素级与图像级指标的定义及适用条件。

2.1 常用数据集

常用数据集提供图像、真伪标签或像素级掩膜,其标注粒度及获取方式决定了可支持的训练与评测任务。按照第 1.2 节的 3 类伪造源,传统局部篡改数据集既包括手工编辑的小样本库,也包括自动合成的大规模库;生成式局部编辑数据集常提供编辑前后的成对图像;本节所列全合成图像数据集则主要提供图像级标签。不同数据集的容量相差 4 个数量级,规模差异也是影响性能比较的因素。IMDL-BenCo 指出,在几百到几千张图像的数据规模下,训练集与预训练数据的差异可能影响模型排名。

传统局部篡改数据集中的手工编辑样本具有不同的真实性与处理条件。哥伦比亚大学的 Columbia 拼接库含 183 张真图与 180 张拼接图,拼接区域未经后处理,并采用无压缩格式保存,因而视觉异常较明显^[47]。中国科学院自动化研究所的 CASIA 库分为两版: v1 的篡改区未经后处理, v2 在拼接或复制粘贴后施加模糊、缩放与旋转,更接近实际编辑条件^[46]。面向复制粘贴检测, Wen 等^[57]的 COVERAGE 在同幅图像中加入相似的真实物体作为干扰, Tralić 等^[65]的 CoMoFoD 则结合旋转、缩放、加噪、模糊与压缩,提供多种退化条件。De Carvalho 等^[66]基于自行采集的室内外照片,通过添加人物并按需调整颜色与亮度构建 DSO-1; 图像保存为 PNG,但多数事先经历过 JPEG 压缩。美国国家标准与技术研究院自 2016 年起发布的 NIST 系列为同一伪造内容提供多个版本,通过不同边界后处理与压缩质量因子考察性能变化^[48]。Zampoglou 等^[67]的 Wild Web 从网络收集实际伪造案例,通过同一案例的多版本比对获得掩膜,并另附可获得的来源真图。来源真图用于追溯伪造过程,不等同于独立配置的评测负样本。

随着深度模型对训练规模的需求增加,自动合成为扩充训练样本提供了便利。Mahfoudi等^[50]基于MS-COCO自动生成DEFACTO,涵盖拼接、复制粘贴、移除与人脸变形,包含22.9万张伪图。Kwon等^[6]随CAT-Net发布的tampCOCO等3个合成库共有60万张伪图,主要用于训练。

社区收集与人工编辑的数据集补充了自动合成样本的操作类型和处理条件。Heller等^[68]的PS-Battles从图像处理爱好者社区收集逾1万组实例,每组包含原图与若干篡改图,但不提供像素级标签。Kniaz等^[69]的Fantastic Reality同时包含未经细致修饰和经过人为润饰的拼接样本,两类各占约一半,并提供物体类别标签。Novozamsky等^[49]的IMD2020包含合成与网络收集两个子集:合成子集以2322种相机型号拍摄的原图为基础,通过拼接、修复与形变等操作生成样本;网络子集收集实际伪造图及对应真图,并由人工标注篡改区域。Korus等^[70]的真实篡改库包含220组图像,来自4台确定型号的相机,并附带模式噪声信息,可用于相机指纹方法的研究。

生成式局部编辑数据集的共同特点是通过生成模型合成编辑区域,而非直接复制已有像素。Guillaro等^[7]随TruFor发布的CocoGlide使用GLIDE扩散模型替换COCO图像中的物体,包含512张伪图及对应的512张真图,分辨率统一为256像素 $\times$ 256像素。Jia等^[51]的AutoSplice通过文本提示进行局部与整体编辑,包含2273张真图与3621张伪图,并为每张图像提供3档JPEG压缩版本,用于比较不同压缩条件。Chen等^[52]的GIM采用自动化流程,由分割模型生成候选掩膜、大语言模型生成替换提示词,再由扩散模型重绘或移除内容,共产生114万对图像,官方评测子集还包含3种随机退化。Xu等^[53]随ManipShield构建的ManipBench覆盖25个编辑模型与12类篡改操作,逾45万张伪图中有10万张附边界框、判据线索与文字解释标注。真伪配对图像用于训练,来源真图总量不代表测试集的负样本数量;原文注明数据集将随论文发表公开。

全合成图像数据集主要提供图像级标签,用于评测整幅生成图像的鉴别能力。与局部编辑数据集相比,这类标签不区分图像内部的区域来源。Zhu等^[54]的GenImage采用8个生成对抗网络(Generative adversarial network, GAN)与扩散生成器,覆盖ImageNet全部类别,真伪图像各逾100万张。官方协议除跨生成器泛化外,还设置退化图像分类任务,评测低分辨率、模糊与压缩3种条件下的判别能力。Wang等^[32]随扩散反演检测方法DIRE发布的DiffusionForensics涵盖由多种扩散模型生成的卧室、通用物体与人脸图像。正式版按领域及生成器列出训练、验证与测试子集,其中部分生成器样本仅用于跨生成器测试。由于缺少局部篡改标注,这类数据集不能直接用于比较区域定位能力。

3类数据集在规模、真图负样本、标注粒度与退化覆盖方面存在差异,表2汇总其构成与出处。手工篡改库多为小样本,自动合成库规模较大;生成式局部编辑库支持掩膜或边界框定位,全合成库则主要支持图像级判别。部分文献对真图负样本的记载不一,编辑工具与后处理条件的覆盖也有差异,跨库比较应据实际使用的子集与标注确定评测范围。

表2 常用数据集全景
Table 2 Overview of commonly used datasets

数据集	规模(真图/伪图)	伪造源类型	像素级 掩膜	真图负 样本	分辨率/ (像素 $\times$ 像素)
Columbia ^[47]	183/180	拼接	有	有	约757 $\times$ 568
CASIA v1/v2 ^[46]	v1为800/921;v2为7200与 5123、7491与5123、7491 与5063(见表注)	拼接、复制粘贴、移除,含 模糊、缩放等后处理	有	各文献 记载不 一	384 $\times$ 256至 800 $\times$ 600

续表

数据集	规模(真图/伪图)	伪造源类型	像素级掩膜	真图负样本	分辨率/(像素×像素)
COVERAGE ^[57]	100/100	复制粘贴,含相似真实物体干扰	有	有	均值400×486
CoMoFoD ^[65]	260/260	复制粘贴,伴随旋转、缩放、加噪、模糊与压缩	有	有	512×512与3 000×2 000
DSO-1 ^[66]	100/100	拼接(含光照一致性判据)	有	有	2 048×1 536
Wild Web ^[67]	90个子案例,9 657张伪图	网络实际伪造	有	附来源真图	不统一
PS-Battles ^[68]	1.1万/10.2万	社区多种编辑操作	无	有	不统一
Fantastic Reality ^[69]	1.6万/1.9万	拼接,半数粗糙半数润饰	有	有	未统一
真实篡改库 ^[70]	220/220	拼接、复制粘贴,附相机模式噪声	有	有	1 920×1 080
NIST16 ^[48]	560/564或0/564(见表注)	拼接、复制粘贴、移除	有	各文献记载不一	500×500至5 616×3 744
IMD2020 ^[49]	合成3.5万/3.5万;网络约2 000对	合成编辑、网络伪造	有	有	均值1 062×866
DEFACTO ^[50]	0(见表注)/22.9万	MS-COCO上自动生成的拼接、复制粘贴、移除、人脸变形	有	见表注	未报告
tampCOCO等 ^[6]	0(见表注)/60万	自动合成的拼接与复制粘贴	有	见表注	未报告
AutoSplice ^[51]	2 273/3 621,每张另有3档压缩版本	文本提示驱动的局部与整体编辑	有	有	256×256至4 232×4 232
CocoGlide ^[7]	512/512	扩散图像修复替换物体	有	有	256×256
GIM ^[52]	114万/114万(成对)	扩散图像修复重绘与扩散零空间模型移除	有	有	64×64至6 000×3 904
ManipBench ^[53]	2.2万来源真图/逾45万伪图	25个模型、12类操作	未报掩膜;10万框标注	训练含真图	未公开报告
GenImage ^[54]	逾100万/逾100万	8个GAN与扩散生成器整幅生成	无	有	未统一
DiffusionForensics ^[32]	13.4万/48.12万	多种扩散模型整幅生成	无	有	未统一

注:除注明外,数据取自各数据集原文;Wild Web采用所引版本,DiffusionForensics按正式版分项加总,不重复计入重建图和残差图。CASIA、NIST16的样本构成在各文献中记载不一;NIST16伪图均记564,真图GIM^[52]记560、MVSS-Net^[5]记0。CASIA v1的真图在MVSS-Net中记为800张,IMDL-BenCo^[55]则因其所用版本缺少真图而未将其纳入图像级评测,二者不能混同。DEFACTO、tampCOCO的真图数据GIM记为0,不表示数据集确无真图;CocoGlide原论文包含配对真图,GIM汇总记为0。

2.2 评价指标

篡改定位在数学形式上是逐像素的二分类,因此沿用分类任务的评价体系:把预测结果按真值分为4类并记其数量为真阳性(True positive, TP)、真阴性(True negative, TN)、假阳性(False positive, FP)与假阴性(False negative, FN)。准确率(Accuracy, ACC)、 F_1 、交并比(Intersection over union, IoU)与马修斯相关系数(Matthews correlation coefficient, MCC)由这些计数导出,受试者工作特征曲线下面积(Area under the curve, AUC)则进一步考察阈值变化时的排序表现。以下逐个给出定义并讨论适用边界。

(1)ACC。ACC定义为判定正确的像素占全部像素的比例,即真阳性与真阴性之和除以4个计数之和。在类别失衡时,这一指标可能掩盖定位误差:它对每个像素等权计数,因此数量较多的原始像素对结果的贡献更大。一幅图被改动的区域通常只占百分之几,把全部像素判为原始即可获得九成以上的准确率,这与模型是否找到篡改区无关。

(2) F_1 分数。 F_1 分数取精确率 Precision=TP/(TP+FP)与召回率 Recall=TP/(TP+FN)的调和平均,同时考虑误报与漏报,相比准确率更能反映少数类的识别情况。 F_1 的解释还受到评测范围与阈值选取的影响。若仅在伪造图像上计算并平均像素级 F_1 ,完整真实图像上的假阳性不计入结果,因而可能忽略模型对真图的误报;MVSS-Net的比较实验^[5]中,GSR-Net^[71]的像素级结果与图像级特异度之间的差异就与此有关。MVSS-Net原文的图像级“ F_1 ”取灵敏度 TP/(TP+FN)与特异度 TN/(TN+FP)的调和平均,与精确率、召回率构成的标准 F_1 定义不同。部分工作对每张概率图选择最优阈值,再计算平均 F_1 ,但实际取证时缺少真值,无法逐图确定最优阈值,因此这类结果不能直接代表部署性能。标准 F_1 分数的定义为

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

(3)AUC。AUC定义为受试者工作特征曲线下的面积,该曲线以假阳性率为横轴、真阳性率为纵轴,通过改变判定阈值得到。它可解释为随机各取一个篡改像素和原始像素时,前者得分高于后者的概率,同分情况计半。AUC取值介于0与1之间,数值越大表示排序能力越强。该指标不依赖单一阈值,但也不直接衡量二值化后的区域重合程度。两个AUC相近的方法在固定阈值下的 F_1 可相差1个量级,因此较高的AUC不必然对应较好的掩膜质量,IMDL-BenCo也指出了单独采用该指标的局限。

(4)IoU。IoU定义为预测篡改区与真实篡改区的交集面积除以并集面积,直接度量区域重合程度,是语义分割领域的通用指标。它与 F_1 在同一测试集上总体正相关,但数值变化尺度不同;预测区域偏移或外扩会增加并集中的非重叠部分,从而降低交并比。同一方法的IoU通常低于其 F_1 ,两种指标不宜直接混用于跨文献比较。IoU的定义为

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$$

(5)MCC。MCC系数以真阳性与真阴性之积减去假阳性与假阴性之积为分子,以4个边缘计数乘积的算术平方根为分母,取值介于-1与1之间,0表示预测与真值之间无线性相关。该指标同时考虑正负两类的判定结果,反映固定阈值下的表现;AUC则衡量跨阈值的排序能力,两者在退化图像上的变化幅度可能不同。MCC的定义为

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

3 鉴伪方法

按主要研究目标与技术侧重点,鉴伪方法分为传统取证线索、生成式伪造分析与多模态大模型3条研究路线,分别侧重成像及编辑痕迹、生成过程相关特征、取证信号与语言推理的结合。三者存在交叉,兼具多种功能的方法按主要贡献归类。

3.1 基于传统取证线索的方法

基于传统取证线索的方法主要利用成像、编码与编辑留下的统计特征。区域一致性方法比较篡改区与背景的处理历史,重采样、复制及语义异常检测则不一定要求图像中存在两种相机来源。噪声域与相机指纹、频域与压缩痕迹、边界与语义不一致构成主要分析线索,可信度建模与主干架构研究分别关注输出可靠性和特征提取方式。不同线索在编辑与后处理中的保留程度各异,其退化顺序也随处理流程变化。

3.1.1 噪声域与相机指纹

成像处理会留下与器材及处理历史有关的统计特征,噪声域方法通过残差分析区域间的不一致。像素响应非均匀性的估计需要同一设备的多张对齐图像^[62],网络图像通常难以满足这一条件。因此,相关方法转向手工残差统计、学习型指纹及元数据自一致性学习,以减少对参考图像或篡改标注的依赖。

Fridrich等^[2]构造的富模型使用多方向高通滤波器抑制图像内容,对邻域预测残差量化、截断后统计高阶共生矩阵,再拼接各子模型的直方图。完整特征达到34 671维,通过冗余子模型覆盖不同残差关系。其描述不限定篡改类型,但固定的滤波方向与阶数仍限制了表达范围。

Bayar等^[72]设计的约束卷积将首层卷积核中心系数固定-1、其余系数之和约束为1,使输出对应邻域预测残差。每次更新参数后重新施加约束,滤波器因而可在抑制内容的条件下由数据学习。该结构被用于多种定位网络的噪声分支,但约束仅作用于首层,后续层仍可能引入语义信息。

Cozzolino等^[19]提出的Noiseprint通过孪生网络学习相机模型共有的指纹,减少对逐台设备标定的依赖。训练时,将来自相同相机模型且位于图像相同位置的图像块组成正样本对,模型或位置不同的图像块组成负样本对。位置约束有助于保留相机内部处理形成的周期性模式,避免不同位置的残差相互抵消。网络输出同尺寸残差图,增强相机模型特征并抑制场景纹理。推断仅需一幅待检图像,指纹图的区域性异常可用于发现来源不一致或空间错位的位置。Comprint^[21]进一步以压缩历史为学习目标,使用孪生网络提取压缩指纹,并与相机指纹融合。

图像元数据也可用于构造自一致性学习任务,从而减少对篡改标注的依赖。Huh等^[73]以一对图像块的可交换图像文件格式(EXIF)标签是否一致作为代理任务,训练网络判断两块是否具有相同的拍摄来源与设置,再将成对预测聚合为全图一致性热图。Mayer等^[74]提出的取证相似度将这一思路扩展到其他取证特征,通过学习图像块是否同源,使模型能够用于未见过的相机型号。此类方法估计的是来源一致性,它与图像真实性并不等价;统一后处理还可能削弱区域差异,具体影响随处理参数变化。

Cozzolino等^[3]提出的Splicebuster根据共生残差特征进行无监督聚类,生成拼接热图。Zhou等^[4]提出的RGB-N在两阶段目标检测框架中,以双线性池化融合富模型残差与RGB特征,输出篡改边界框。BusterNet^[20]则分别检测图内相似区域与可疑篡改区域,再融合两分支结果,区分复制粘贴的源区和目标区。

噪声域方法的主要依据是残差中的统计特征,区域一致性检测尤其依赖图内差异的可分辨性。再压缩、缩放与生成式重绘可能削弱这些差异,其影响与操作强度、背景保留方式有关。全合成图像没有

经历真实相机成像过程,经典相机指纹缺少对应的物理来源,检测时则可能利用生成器留下的残差特征。

3.1.2 频域与压缩痕迹

JPEG 压缩对分块 DCT 系数进行量化,区域间不同的压缩历史可能形成可检测的统计差异。例如,已压缩的背景与未经压缩的粘贴区再次保存后,可能分别经历两次与一次量化,系数直方图因而呈现不同的周期性特征。

Popescu 等^[1]通过期望最大化估计各像素与邻域像素线性相关的概率,再检测概率图频谱中的周期峰,以识别缩放或旋转痕迹;其能力受重采样方式与保存质量影响。Bianchi 等^[63]为每个 8×8 变换块估计一次或两次量化的后验概率,经过空间平滑得到篡改热图,无需训练。

Kwon 等^[6]提出的 CAT-Net 将压缩伪影学习纳入端到端检测网络。该网络包含 RGB 流与 DCT 流两路并行的高分辨率主干^[75]:RGB 流处理图像内容,DCT 流将逐个 8×8 块中的量化系数截断并取绝对值,编码为保留空间位置的二值 DCT 系数体。压缩伪影学习模块按频率组织卷积运算,并结合量化表提取压缩历史的统计特征;两路特征融合后输出像素级篡改概率图。对于非 JPEG 图像,网络从 RGB 像素计算 DCT,并采用全一量化表,因此最终以无损格式保存的图像仍可由该网络处理。DCT 支路的作用取决于可利用的 JPEG 压缩痕迹,再压缩可能削弱区域间的统计差异,RGB 支路则提供图像内容线索。Guo 等^[22]提出的 EITLNet 将图像与高通残差分别送入双分支编码器,并通过共注意力在多个尺度上融合,使分割主干的同时可利用图像内容和高频特征。

双压缩分析依赖 JPEG 量化历史的可分性,一般频域特征则不限于 JPEG 痕迹。强再压缩可能削弱区域差异,数据增广只能改善对残余线索的利用,不能保证恢复已丢失的信息。因此,保存格式、压缩参数与训练设置共同影响性能。

3.1.3 边界与语义不一致

拼接可能造成光照、阴影、透视关系或边界高频特征的不一致。图内异常检测比较局部特征与整体分布,边界监督则直接学习篡改轮廓。部分几何与语义异常在压缩或降采样削弱细节后仍可能保留。

图内异常检测通常不针对具体篡改类型建模。Wu 等^[23]提出的 ManTra-Net 先以 385 类图像操作的分类任务预训练痕迹提取器,使其重点表征操作痕迹并抑制语义内容。检测时,循环单元逐步累积全图统计量,将局部特征与全图分布比较,生成表示偏离程度的异常响应图。该方法无需针对目标数据集微调,但输出是未经概率校准的异常响应,其强弱还与篡改区在图内所占的比例有关。

边界监督以篡改掩膜提取的轮廓作为训练信号,Chen 等^[5]提出的 MVSS-Net 将这一信号与噪声特征结合。网络的边缘监督分支在多层特征图上施加 Sobel 算子以增强边缘响应,再逐级融合特征并预测篡改边界;噪声分支以约束卷积提取低层残差,补充边界之外的统计信息。两路特征通过包含通道注意力与位置注意力的双注意力模块融合,输出像素级掩膜,并经全局最大池化得到图像级判定。训练时还引入真实图像作负样本以降低误报,期刊扩展版延续了多视角与多尺度监督的设计^[76]。GSR-Net 将边界信息用于样本生成与掩膜优化,通过生成、分割、细化 3 个阶段逐步改善预测边界。

Hu 等^[77]提出的 SPAN 以局部自注意力金字塔逐级增大采样间距,在保持特征分辨率的同时扩大比较范围。Liu 等^[56]提出的 PSCC-Net 先生成低分辨率掩膜,再通过空间与通道相关模块逐级上采样、细化。四级掩膜在同一次前向计算中生成,同时输出图像级分数。粗尺度结果为细尺度提供先验,但漏检也可能传递至后续阶段。

ObjectFormer^[24]从 RGB 图像及高频特征生成图像块表示,以交叉注意力更新可学习的对象原型,

再将聚合的上下文传回图像块。边界敏感模块计算像素与局部邻域特征的余弦相似度,突出篡改边界。分类头与分割头分别输出图像级判定和像素级掩膜,将对象级上下文与局部一致性分析结合。

Niloy 等^[78]提出的 CFL-Net 通过对比学习缩小篡改像素间的特征距离、增大篡改区与真实区的距离,不预先限定判别线索。Hao 等^[79]提出的 TransForensics 通过编码器各层间的密集自注意力融合浅层纹理、残差与深层语义,减少下采样造成的痕迹损失。

边界监督依赖局部轮廓差异,语义分析还可利用内容与几何异常。生成式编辑及融合可能减弱边界响应,大面积重绘可能破坏图内异常检测所需的背景参照,具体影响随方法与编辑流程变化。全图合成时,真值掩膜为全“1”矩阵,图内缺少可供比较的真实区域,也没有真实与篡改区域之间的边界。

3.1.4 可信度建模与架构演进

定位掩膜并不能充分反映预测的可靠性,置信度建模因而成为独立的研究问题。架构研究则关注主干网络设计及取证先验的融合方式,主干与先验各自的贡献需通过消融分析区分。

Guillaro 等^[7]提出的 TruFor 同时预测篡改位置与局部置信度。该方法先以包含编辑与再压缩操作的真实图像训练 Noiseprint++ 提取器,从待检 RGB 图像中获得对相机内处理与编辑历史敏感的残差图。RGB 图像与残差图分别经过两条 SegFormer^[80] 分支提取多尺度特征,并在各阶段通过跨模态校准与融合交换信息。融合特征输入两个并行解码器,分别生成逐像素篡改概率图和表示判定可信程度的置信图。置信图为异常图的聚合统计量提供权重,以降低低可信区域的影响,分类器再据此得到图像级完整性分数,架构如图 5 所示。

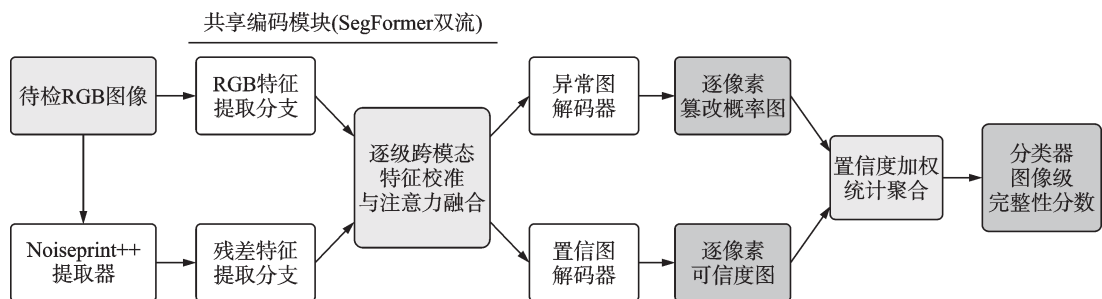


图 5 TruFor^[7]双解码器架构

Fig.5 Dual-decoder architecture of TruFor^[7]

在特征提取架构方面, Ma 等^[25]提出的 IML-ViT 分析了高分辨率输入、多尺度特征与边缘监督 3 项要素的作用。该方法以掩码自编码器^[81]预训练的标准视觉 Transformer^[82]为主干,结合特征金字塔与边缘损失完成定位,不设置噪声或频域分支。Su 等^[26]提出的 SparseViT 同样不依赖人工设计的噪声与频域分支。该方法按预设稀疏率将特征分组,在各组内部独立计算自注意力,以减少连续语义信息的交互并突出非语义篡改痕迹。不同层采用不同稀疏率,预测模块再融合多尺度特征,生成定位掩膜。Zhu 等^[27]提出的 Mesorch 则保留不同频段的先验,采用结合卷积与 Transformer 的“介观”混合架构。该方法先用离散余弦变换将输入分为高低两个频段,由卷积分支提取高频段的压缩与噪声痕迹,由 Transformer 分支提取低频段的语义与结构异常。融合模块依据输入自适应生成权重,以适应不同篡改类型和压缩强度下两类特征重要性的变化。当两类线索同时被削弱或抹除时,这种融合仍会受到限制。

ForensicHub^[10]通过消融实验考察了低层取证先验与不同主干的组合。在将 Sobel、约束卷积、频域

高通与DCT特征以浅层拼接方式加入强主干后,几乎全部组合的平均图像级AUC下降,ConvNeXt^[83]与Swin Transformer^[84]的变化较明显。EfficientNet^[85]与ResNet^[86]等轻量主干的部分组合略有提升,表明该设置下的先验融合效果与主干有关。实验采用统一降采样到256像素×256像素的跨域协议,结论仅适用于浅层拼接,尚不能推广为低层特征提取器普遍无效。实验将4类低层提取器与6种主干两两配对,强主干上的平均图像级AUC下降0.049~0.098,其中ConvNeXt与Swin Transformer的最大降幅分别为-0.0983与-0.0938,轻量主干上的最大增幅为+0.0367^[10]。

IMDL-BenCo^[55]的统一复现结果反映了评测设置对架构比较的影响。该基准设有两套训练协议:Protocol-MVSS仅以CASIAv2训练,Protocol-CAT则使用CAT-Net的混合训练集,并将已用于训练的IMD2020从测试集中排除,两套协议分别报告五库和四库均值。在Protocol-CAT对8个方法的评测中,IML-ViT的平均像素级 F_1 为0.725,CAT-Net为0.601。改用Protocol-MVSS后,两者分别为0.519和0.401,TruFor为0.530;TruFor在这一协议下的像素级AUC与交并比也取得较优结果。从Protocol-MVSS改用Protocol-CAT,IML-ViT的均值增加0.206,TruFor增加0.097,但这些差值同时受到训练数据与测试集构成变化的影响,不能单独解释为训练数据带来的增益。检测与定位结果也存在差异。在带检测头的3个方法中,MVSS-Net的像素级 F_1 为0.341,图像级 F_1 为0.631。同一协议下,PSCC-Net与ObjectFormer的像素级 F_1 分别为0.333与0.281。ManTra-Net的像素级 F_1 为0.123,AUC为0.783,说明其异常响应具有一定区分能力,但准确生成掩膜仍有困难^[55]。

可信度建模与架构改进均受训练数据和测试分布影响。现有复合退化实验尚未独立验证置信图对不可靠区域的预警能力^[87],其可靠性仍需单独评价。不设置手工先验分支的方案,其效果也随训练协议变化。现有比较尚难将性能变化单独归因于先验分支或主干容量,训练分布及退化条件的覆盖仍影响方法的适用范围。表3汇总代表方法的统一复现结果,并注明原文结果无法配对比较的原因。复现变化还因数据集而异,原报告并非普遍虚高:MVSS-Net在CASIAv1上由0.452提高至0.534,在Columbia上则由0.638降至0.386^[55]。

表3 传统取证线索代表方法对比

Table 3 Representative methods based on traditional forensic traces

方法	年份	主线索	F_1 (MVSS 协议)	F_1 (CAT 协议)	CASIAv1 原文→复现	主要局限
Splicebuster ^[3]	2015	局部共生噪声特征	未纳入	未纳入	未报掩膜级指标	同机拼接、全图重绘
RGB-N ^[4]	2018	富模型噪声流加RGB流	未纳入	未纳入	仅报逐图最优阈值 F_1	只输出边界框、重压缩敏感
BusterNet ^[20]	2018	相似区域与篡改区域并行	未纳入	未纳入	仅报复制粘贴子任务	拼接与重绘、源区被裁去
RRU-Net ^[88]	2019	环形残差强化边界响应	未纳入	未纳入	训练设置不同	边界被平滑、大面积重绘
ManTra-Net ^[23]	2019	操作痕迹的图内异常	0.123	0.295	0.155→0.125	大面积篡改、重绘、重压缩
CR-CNN ^[89]	2020	区域提议引导细粒度分割	未纳入	未纳入	原文按数据集分列	分割依赖候选区域

续表

方法	年份	主线索	F_1 (MVSS 协议)	F_1 (CAT 协议)	CASIAv1 原文→复现	主要局限
SPAN ^[77]	2020	多尺度局部自注意力	未纳入	未纳入	按是否微调分列 两组	大幅缩放、跨库不 微调
MVSS-Net ^[5]	2021	Sobel边缘监督加噪声视图	0.341	0.547	0.452→0.534	重绘平滑边界、移 除类无边界
TransForensics ^[79]	2021	各层间密集自注意力	未纳入	未纳入	训练设置不同	单一信号域痕迹 被抹除
CAT-Net ^[6]	2022	二值DCT系数体与量化表	0.401	0.601	训练设置不同	压缩差异弱、强再 压缩
ObjectFormer ^[24]	2022	对象级图像块一致性	0.281	0.447	仅报微调后 AUC	纹理区小面积篡 改、复现困难
PSCC-Net ^[56]	2022	多尺度空间与通道相关	0.333	0.551	训练集非 CA- SIAv2	粗尺度漏检影响 细化
NCL-IML ^[90]	2023	非互斥对比学习	0.334	0.467	0.598→0.502	跨库分布差异大
CFL-Net ^[78]	2023	篡改与真实区特征对比	未纳入	未纳入	训练设置不同	重绘区与背景表 示趋同
TruFor ^[7]	2023	Noiseprint++ 噪声域加置信图	0.530	0.627	用了额外预训练 数据	高斯噪声、复合退 化
IML-ViT ^[25]	2023	高分辨率加边缘监督	0.519	0.725	0.721→0.721	小样本、强降采样 协议
EITLNet ^[22]	2024	高通残差与共注意力融合	未纳入	未纳入	训练设置不同	全图统一再压缩
Mesorch ^[27]	2025	DCT 分频的介观混合	未纳入	未纳入	未参与统一复现	两类线索同时被 抹除
SparseViT ^[26]	2025	稀疏注意力的非语义特征	未纳入	未纳入	未参与统一复现	分布外压缩缺先 验补足

注: F_1 为IMDL-BenCo的像素级均值;MVSS、CAT协议分别指 Protocol-MVSS、Protocol-CAT,对应五库和四库评测;“未纳入”表示未参与该协议评测;原文与复现结果不能配对时列明原因;局限据方法原文及后续研究归纳。

3.2 面向生成式伪造的方法

生成式伪造既包括整幅合成,也包括局部内容编辑,两类对象所保留的取证线索有所不同。整幅合成图像的检测主要利用生成器伪影、预训练特征和重建残差;局部编辑的分析还包括篡改区域定位与来源归因。本节按整幅生成图像检测、生成式局部编辑分析的顺序介绍相关方法,其中整图检测进一步区分直接提取图像特征与利用扩散反演残差的方法。主动取证通过预先嵌入水印或绑定可信凭证提供来源信息,与从图像中提取痕迹的被动取证互为补充。

3.2.1 全合成图像检测

全合成图像的像素均由生成器产生,通常不存在不同成像历史之间的拼接边界,因此难以通过编辑区与真实背景的局部比较完成检测。常用检测特征主要分为低层伪影与语义特征。低层伪影与生成器的实现细节有关:解码器逐级放大低分辨率特征图时,部分转置卷积与插值配置会使生成图像的频谱偏离自然图像。Durall等^[91]对此进行了系统实证;Frank等^[92]进一步指出,这种偏差在离散余弦变

换谱上表现为具有区分性的系统性峰值。语义特征则用于表征图像内容与结构,相关方法借助预训练模型学习真实与生成图像在特征空间中的差异^[93]。两类特征对生成器结构和训练数据的依赖不同,融合方法尝试利用它们的互补性改善跨生成器检测。

CNNDetection^[28]利用低层伪影开展跨生成器检测。该方法使用单一生成器的数据训练卷积分类器,生成图像仅取自 ProGAN^[61],并施加 JPEG 压缩与高斯模糊增广。分类器从训练样本中学习上采样痕迹,输出整幅图像的真伪概率,以减少对新生成器逐一重新训练的需求。原文报告,跨 11 种未见卷积生成器的平均精确率(Average precision, AP)均值约为 0.93。Tan 等^[29]提出的 NPR 将上采样痕迹表示为相邻像素之间的差分。该方法直接以差分结果作为网络输入,无需频域变换或梯度反传。上采样虽作用于特征图,却会影响输出像素的局部相对关系,这一关系构成其检测依据。同一团队提出的 LGrad^[94]采用梯度域表征,FreqNet^[95]则对中间特征进行频域滤波。这些方法分别从局部差分、梯度与频率信息中提取伪影,其迁移效果仍受训练来源与测试生成器差异的影响。

频域方法通过图像变换显式提取生成伪影。Bamney^[64]提出的 Synthbuster 先用高通滤波提取残差,再分析残差的傅里叶谱。扩散模型解码器的上采样会在频谱中形成周期性栅格状峰值,其位置与上采样因子有关,因而可根据特定频点的能量进行检测。Corvi 等^[96]发现,扩散生成图像与 GAN 生成图像具有不同类型的频域指纹,为 GAN 训练的检测器迁移到扩散图像时性能会显著下降。Zhong 等^[97]提出的 PatchCraft 利用局部纹理差异,减少对单一全局统计量的依赖。该方法将图像划分为纹理丰富区与纹理贫乏区,分别提取高通噪声模式,再利用两类区域的像素相关性对比度进行分类。其依据是真实图像中两类区域的像素相关性差异通常大于生成图像。UnivFD^[30]利用预训练语义特征进行检测,以降低对特定上采样算子及其低层伪影的依赖。该方法冻结对比语言图像预训练模型(Contrastive language-image pre-training, CLIP)^[98]的视觉主干,仅在提取的特征上训练线性探针。其有效性依赖预训练特征的通用性,以及真实与生成图像在特征空间中可迁移的差异,无需为每种生成器重新学习表征。Liu 等^[99]提出的 FatFormer 在此类预训练特征中引入伪造感知的自适应模块,对表征进行取证任务微调。Tan 等^[93]将 UnivFD 的有效性与文本侧的类别语义先验联系起来,并在检测过程中显式引入该先验。ForensicHub 的统一复现还考察了语义特征在局部篡改检测中的表现。在 IMD2020^[49]上,UnivFD 的图像级 AUC 为 0.786,4 个 IMDL 专用方法的结果为 0.539~0.570。这一比较反映了该数据集和统一设置下的整图判别能力。

AIDE^[31]融合低层统计与语义特征。该方法通过块级高低频统计选择伪影较明显的图像块,提取局部特征,同时利用 CLIP 获取全局语义特征,再融合两路信息进行检测。局部特征用于保留生成过程的细节痕迹,语义特征用于补充跨生成器的内容与结构信息。该方法发布的 Chameleon 基准包含人工筛选的高保真生成图像,其视觉外观较难与真实图像区分。作者在该基准上测试了包括 AIDE 在内的多个检测器,多种方法出现明显退化,部分接近随机猜测水平。生成伪影还可用于来源归因。Yu 等^[100]通过学习并分析 GAN 指纹,发现不同生成对抗网络的输出图像中存在稳定且可区分的模型级痕迹。这类痕迹为层次化归因方法区分生成器提供了依据。

整幅合成图像的检测通常由特征提取与分类组成,推理开销取决于主干网络 and 处理的图像块数量。低层特征的迁移依赖待检生成器与训练分布共享的上采样或频域伪影,再压缩与缩放可能破坏其依赖的邻域关系。语义特征的迁移则要求真实与生成图像的差异仍能由预训练模型表征。Chameleon 上的退化表明,高保真生成图像可能使既有低层与语义特征同时难以支持准确检测,但这一结果尚不能证明两类图像在所有特征空间中均不可分。

3.2.2 重建与扩散反演残差

重建残差方法利用外部扩散模型分析图像与生成分布的关系。通过近似反演,输入图像被映射到初始噪声,再经采样过程重建;原图与重建图之间的差异用于表征该图像与模型可达流形的距离。与重建模型分布相近的生成图像可能更容易被重建,真实图像则可能呈现不同的残差分布。残差能否用于分类取决于反演误差与模型匹配程度,因此它提供了区别于直接提取低层或语义特征的检测依据。

DIRE^[32]利用扩散反演与重建残差进行整图判别。该方法采用消融扩散模型(Ablated diffusion model, ADM)^[101],对待检图像进行去噪扩散隐式模型(Denoising diffusion implicit model, DDIM)^[102]反演与重建。原图与重建图的逐像素绝对差构成 L_1 距离图,卷积分类器以此为输入学习真伪边界。模型输出整幅图像的判别结果,原文报告在DiffusionForensics上ACC与AP均接近0.99。Chen等^[33]提出的DRCT将扩散重建用于构造困难训练样本。该方法结合对比损失与分类损失学习特征,减少对训练生成器特有伪影的依赖。重建仅用于准备训练样本,推理时由训练后的检测器直接判别待检图像,无需额外反演。

重建残差不要求图像中存在拼接边界,也不要求重建器与待检生成器同类。DIRE原文在固定ADM重建器的条件下,观察到对Stable Diffusion^[58]等不同生成器的迁移能力。生成器、分类器训练分布与重建器分别影响检测结果,残差的可分性仍需在目标数据上验证。重建所处的阶段还决定了推理代价:DIRE在检测时逐图反演与重建,数十步迭代增加了单图计算量;具体开销取决于迭代步数、重建模型与硬件条件,不能由生成器类型直接推定。

3.2.3 生成式局部编辑的定位与归因

生成式局部编辑的分析需要同时识别图像真伪与编辑区域。重绘和区域融合可能削弱传统篡改痕迹,背景残差与边界差异也随编辑流程而变化,因而需要相应的像素标注数据进行训练和评测。TruFor^[7]发布的CocoGlide采用GLIDE^[59]重绘构造,规模较小,主要用于检验既有方法在扩散重绘上的性能。GIM^[52]以通用分割模型SAM^[103]生成掩膜,再由扩散零空间模型DDNM^[60]完成图像修复,提供百万量级的成对训练与评测样本。ManipBench^[53]与LocateEdit-Bench^[104]提供指令式编辑的定位标注,覆盖近期的多模态编辑模型。

像素不一致性建模(Pixel-inconsistency modeling, PIM)^[34]的跨库直测比较了不同编辑流程下的定位性能。该实验采用像素级MCC,测试集包含AutoSplice的3档压缩版本、CocoGlide等公开数据集。利用压缩线索的CAT-Net^[6]在AutoSplice^[51]质量因子100档上取得0.578,在CocoGlide上为0.133。两项结果同时涉及数据集与编辑过程的差异,无法单独确定退化原因;面向传统局部篡改设计的方法在生成式编辑上的有效性,也不能由传统篡改结果直接推定。

扩散去噪任务可为局部编辑定位提供预训练表征。Yu等^[105]提出的DiffForensics采用两阶段训练,先固定经通用分割任务预训练的编码器,以真实图像和伪造图像进行去噪扩散预训练,再联合微调编码器与解码器,学习真伪分类、掩膜预测和边缘定位。其测试既包括传统篡改,也包括AutoSplice^[51]等生成式编辑;多数比较方法使用相同训练数据重训,另有部分方法直接采用公开权重。Wang等^[35]提出的ForgDiffuser将定位表述为条件生成任务。该方法以待检图像为条件,通过扩散模型逐步去噪并采样得到篡改掩膜,区别于直接回归概率图的定位方式。其评测覆盖传统篡改与生成式编辑,各数据集分别划分训练集和测试集。Nam等^[106]提出的M2SFormer侧重边界质量,在跳跃连接处引入多光谱多尺度注意力,并通过边缘难度引导网络加强对难分边界的建模。该方法在拼接、复制粘贴等通用篡改数据上进行跨库评测,尚未给出专门的扩散重绘测试结果。

来源归因用于识别伪造类型或生成模型,可与区域定位并行开展,为司法取证与平台治理提供来源信息。HiFi-Net^[8]采用4层分类树,先区分全合成与局部篡改,再逐级细分至具体生成器。网络同时输出定位掩膜,在同一次前向计算中完成真伪判别、区域定位与来源归因。其期刊扩展版 HiFi-Net++^[107]利用语言先验引导特征对齐,以改善对未见类别的迁移,减少对训练数据枚举全部类别的依赖。Sha等^[108]提出的 DE-FAKE 面向文生图场景,在判断真伪的同时识别生成图像的文生图模型。实验表明,配对文本提示可同时改善其判别与归因性能。VLForgery^[109]利用多模态大模型联合输出真伪结果、定位掩膜与来源归因;ManipShield^[53]则从多模态大模型的内部特征中提取检测依据,输出真伪判定、编辑区域边界框与文字解释,其中的取证线索类别描述可疑现象,不对应编辑模型身份。

DiffForensics、ForgDiffuser 与 M2SFormer 分别从预训练表征、掩膜生成和边界建模改善定位,实验所覆盖的伪造类型、数据划分与基线训练方式有所不同。跨库测试反映向未见数据的迁移,同库划分则主要检验相近分布下的定位性能。依赖物体结构的掩膜预测还受到编辑范围的影响:完整物体可以提供结构参照,不完整的编辑对象则可能与物体轮廓不一致。来源归因受类别范围限制,待检生成器位于预定义类别集合内时,模型能够提供比真伪标签更细的来源信息;新生成器缺少对应类别时,封闭集分类器仍会将其分配至已有类别。ForensicHub 的统一跨域评测中,HiFi-Net 在 11 个数据集上的平均图像级 AUC 为 0.668,在 22 个受测方法中列第 18^[10],显示出跨域检测的局限。上采样算子与后处理流程会影响模型指纹,解码器变化也可能改变归因特征,因此来源类别扩展与特征变化均会影响归因的有效性。PIM 的跨库评测还比较了传统定位器与像素不一致性模型在扩散重绘图像上的表现,表 4 列出其中代表方法在 CocoGlide 上的像素级 AUC 与 MCC^[34]。

表 4 代表方法在 CocoGlide 上的像素级定位性能

Table 4 Pixel-level localization performance of representative methods on CocoGlide

方法	AUC	MCC
ManTra-Net ^[23]	0.806	0.190
SPAN ^[77]	0.776	0.198
PSCC-Net ^[56]	0.566	0.051
MVSS-Net++ ^[76]	0.819	0.309
CAT-Net ^[6]	0.587	0.133
TruFor ^[7]	0.757	0.253
PIM ^[34]	0.817	0.372

注:数据取自 PIM 跨库评测;CocoGlide 未参与训练, MCC 阈值为 0.5。

3.2.4 主动取证的互补

主动取证在内容生成或发布时预先记录可验证的来源信息,通过验证这些信息的完整性追踪内容来源。它依赖生产或发布环节的配合,与从像素中分析伪造痕迹的被动取证形成互补。按信息的附着位置,相关技术包括文件容器中的来源凭证、生成模型权重中嵌入的签名,以及采样起点中嵌入的水印;不同位置对应不同的鲁棒性与部署条件。

内容来源与真实性联盟(Coalition for content provenance and authenticity, C2PA)的内容凭证^[36]将来源信息绑定至文件容器。依据其 2.4 版规范,设备信息与编辑历史被组织为带数字签名的清单,并与资产文件绑定。验证方通过核查签名链的完整性追踪已记录的采集与编辑过程,验证结果不直接判定内容真伪。Fernandez 等^[110]提出的 Stable Signature 将不可见签名嵌入生成图像。该方法微调潜空间扩散模型的解码器,使生成图像携带二进制签名,再由固定的解码网络提取比特串。来源信息随像素传播,因而不依赖文件容器保存。Zhang 等^[37]提出的 EditGuard 同时嵌入版权水印与定位水印,验证端分别提取归属信息与篡改掩膜,将主动取证扩展至区域定位。Tree-Ring 水印^[38]在采样起点嵌入来源信息。该方法在初始噪声的傅里叶域加入同心环图案,使水印随去噪过程分布到生成图像中。验证时通过反演恢复噪声域表征,再检测其中的环形结构。

3类技术均需要相关参与方配合部署,验证代价则有所不同:来源凭证与比特签名主要涉及校验或提取,基于反演的水印还需执行生成模型的逆向过程。截图与平台后处理可能剥离容器级清单,解码器换回原始版本后,权重级水印也随之失效。Tree-Ring需要模型访问权才能完成反演,其验证过程也受到访问条件的限制。主动取证仅覆盖已记录来源信息的内容;凭证缺失时无法据此判断真伪,既不能将无凭证图像认定为伪造,也不能认定为真实。因此,凭证校验可与被动检测结合,分析无有效凭证图像的像素痕迹^[111]。

生成式伪造分析的依据由图像中的生成器伪影,扩展至预训练表征、重建残差及预先记录的来源信息。反演方法的有效性取决于重建模型与目标分布的关系,来源归因受可识别类别范围限制,主动取证则依赖生成或发布环节的配合。虽然部分方法降低了对拼接边界的依赖,重绘和后处理仍可能改变其检测特征;局部编辑定位也可能依赖背景与编辑区之间的差异。表5汇总了各方法原文及后续复现研究中的主要特点与适用条件。

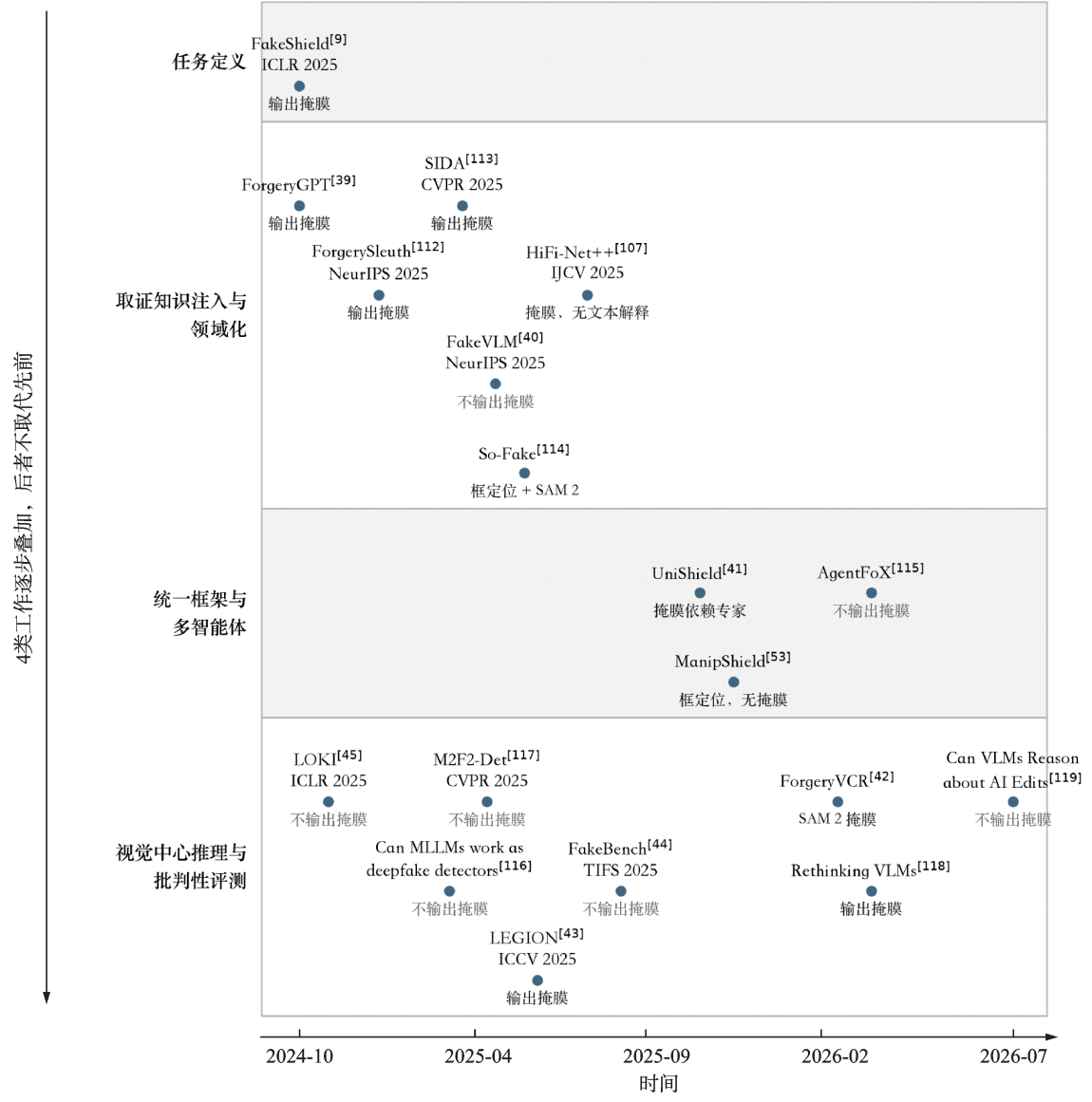
表5 面向生成式图像伪造检测的代表性方法对比
Table 5 Comparison of representative methods for generative image forgery detection

方法	年份	检测对象	线索类型	所需先验	输出掩膜	主要局限
CNNDetection ^[28]	2020	全合成图像	卷积上采样 低层伪影	单一生成器数据 加强增广	否	迁移到扩散生成器后退化
UnivFD ^[30]	2023	全合成图像	CLIP冻结语义 特征	大规模预训练视觉- 语言模型	否	受预训练特征迁移能力限制
DIRE ^[32]	2023	全合成图像	扩散反演 重建残差	可访问的扩散 重建模型	否	目标分布下残差不可分
HiFi-Net ^[8]	2023	局部编辑与 全合成	多分支频域特征 加层次分类树	预定义生成器分类树	是	树外新生成器无法归因、跨域退化
Stable Signature ^[110]	2023	生成端溯源	解码器内嵌 不可见水印	生成方微调解码器	否	解码器换回原版即失效
Tree-Ring ^[38]	2023	生成端溯源	初始噪声傅里叶 域环形水印	生成方配合与 模型访问权	否	验证需模型访问权、再生成可擦除
Synthbuster ^[64]	2024	全合成图像	高频谱域周期 栅格	无	否	质量因子降至90后衰减
NPR ^[29]	2024	全合成图像	邻域像素 关系差分	无	否	再压缩与缩放破坏邻域关系
DiffForensics ^[105]	2024	传统篡改与 生成式编辑	扩散去噪预 训练特征	通用分割预训练权重	是	先验域不匹配时退化
EditGuard ^[37]	2024	版权溯源与 篡改定位	版权水印加定位 水印双嵌入	发布前嵌入水印	是	严重全图篡改影响水印恢复
AIDE ^[31]	2025	全合成图像	块级高低频统计 加CLIP语义	大规模预训练 视觉-语言模型	否	Chameleon上检测性能下降

注:年份按所引文献;线索与先验据方法原文整理,局限结合原文实验与后续研究归纳。

3.3 多模态大模型驱动的可解释方法

检测分数、定位结果与来源标签可以描述取证结果,但通常不能直接说明判别理由。多模态大语言模型(Multimodal large language model, MLLM),文中简称多模态大模型,将视觉编码器的图像特征投影到语言模型的词元空间,由同一个自回归解码器处理图文输入。在取证任务中,这类模型能够生成自然语言解释,描述图像中的异常位置及其判别依据。相关研究从含自然语言解释的任务定义逐步扩展到取证知识注入、跨域协作与视觉中心推理,演进脉络如图6所示,代表工作覆盖至2026年。语言先验也可直接用于定位特征的学习;相关方法的输出能力和解释质量评测方式如表6所示。



注: VLM为视觉语言模型; MLLM为多模态大语言模型。CVPR、ICCV、ICLR、NeurIPS为会议简称, IJCV、TIFS为期刊简称; 节点文献标注与参考文献对应。So-Fake、ManipShield按初稿发布时点标示, 能力按所引版本整理。

图6 多模态与语言引导IFDL方法的演进时间轴
Fig.6 Evolution of multimodal and language-guided methods for IFDL

表 6 多模态与语言引导 IFDL 方法对比

Table 6 Comparison of multimodal and language-guided methods for IFDL

方法或基准	年份	架构范式	像素级 掩膜	文本 判据	类型或生 成器归因	解释质量 评测方式	配套数据集 或基准
FakeShield ^[9]	2025	域标签检测+ 文本引导分割	有	有	域级三分	大模型打分	MMTD-Set
ForgeryGPT ^[39]	2024	掩膜感知提取器 前置串联	有(前置 提取器)	有	—	文本一致性 指标+人工评价	—
ForgerySleuth ^[112]	2025	语义与痕迹 解码融合	有	有	—	语义相似度+ 大模型评分	ForgeryAnaly- sis
SIDA ^[113]	2025	分割词元外接 解码器	有	有	全合成与 局部篡改	大模型打分	SID-Set
So-Fake ^[114]	2025	多路证据门控 与决策	外接 SAM 2	有	篡改与全 合成	文本重合与 语义相似度	So-Fake-Set / OOD
FakeVLM ^[40]	2025	纯图文生成, 无分割头	无	有	—	文本重合与 语义相似度	FakeClue
HiFi-Net++ ^[107]	2025	CLIP类别 文本引导	有	无	分层分类 树四级	不适用	HiFi-IFDL
UniShield ^[41]	2025	双智能体路由 至专家	依赖所选 专家	有	依赖所选 专家	—	—
ManipShield ^[53]	2025	隐层特征+ 任务解码器	无 (框定位)	有	—	文本嵌入余弦 相似度	ManipBench
AgentFoX ^[115]	2026	智能体编排 融合多专家	无	有	—	—	—
M2F2-Det ^[117]	2025	检测器特征桥 接入大模型	无	有	—	与人工解释 标注比对	—
LEGION ^[43]	2025	CLIP与SAM 双编码器	有(局部 伪影)	有	—	文本重合与 语义相似度	SynthScars
ForgeryVCR ^[42]	2026	工具视觉推理+ 分割细化	有 (SAM 2)	主配 置无	—	未单独评测 文本质量	—
FakeBench ^[44]	2025	检测/推理/ 解释三层设问	无	—	伪影类别	人工评分+ 大模型打分	FakeBench
LOKI ^[45]	2025	五模态粗细 两级设问	无	—	伪影类别	人工标注选择题	LOKI

注:“—”表示未见公开报告。FakeBench、LOKI为评测基准,“架构范式”列表示设问方式。OOD表示分布外测试集。So-Fake、ManipShield年份取初稿,能力按所引版本整理。

3.3.1 任务提出与取证知识注入

FakeShield^[9]提出了可解释IFDL(Explainable IFDL, e-IFDL),将真伪判定、篡改掩膜与自然语言解释纳入同一任务。如图7所示,该框架采用两阶段级联结构,将检测与解释的结果用于引导定位。第1阶段的域标签引导可解释检测模块(Domain tag-guided explainable forgery detection module, DTE-FDM)通过视觉编码器提取图像特征,再由大模型生成真伪判定与解释文本;域标签区分软件篡

改、Deepfake与人工智能生成或编辑3类来源,以缓解多源混合训练时监督信号的冲突。第2阶段的多模态取证定位模块(Multimodal forgery localization module, MFLM)将解释文本作为分割提示,利用文本引导的分割解码器生成像素级掩膜。配套的多模态篡改描述数据集MMTD-Set由GPT-4o辅助标注解释文本,为两阶段训练提供图像、掩膜与判据三元组。围绕解释生成所需的取证知识,相关方法将取证信号与语言表征结合,或利用伪影描述提供训练监督。

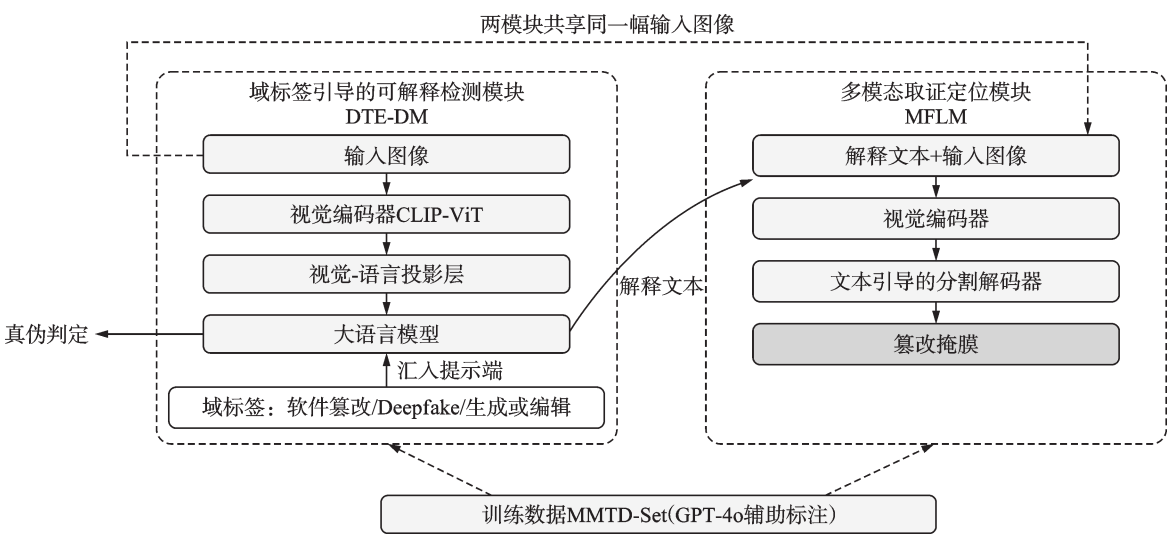


图7 FakeShield^[9]的两阶段架构

Fig.7 Two-stage architecture of FakeShield^[9]

将取证信号转换为词元,可以为解释生成提供局部图像信息。ForgeryGPT^[39]采用掩膜感知特征提取器,将像素级掩膜编码为词元并输入大模型。训练依次经历图文对齐预训练、掩膜与文本对齐预训练、任务指令微调3个阶段。前两阶段分别将整体图像特征、局部掩膜与语言表征对齐,最后通过取证任务对话微调,建立解释文本与空间区域的对应关系。Sun等^[112]提出的ForgerySleuth将大模型的语义分析与独立痕迹编码器结合,以同时利用高层异常与低层统计信息。痕迹编码器通过约束卷积抑制图像内容,提取篡改相关特征。视觉解码器融合痕迹特征、视觉内容特征及大模型的分割词元表征,生成掩膜;语言头输出推理与证据说明。其ForgeryAnalysis语料由GPT-4o生成初稿,再经专家修订,自动数据引擎用于扩展预训练样本。高层解释与低层定位由不同分支处理,解释是否准确反映低层痕迹仍需单独验证。

面向社交媒体常见的全合成与局部篡改内容,SIDA构建了图像检测数据集SID-Set,通过分割词元引导外接解码器,联合优化检测、定位与解释任务^[113]。So-Fake^[114]进一步扩充内容与生成工具的覆盖范围,并以社交平台真图和未参与训练的生成器构建分布外测试集。其配套模型So-Fake-R1由视觉观察器提供语义线索,取证证据模块提取低层异常并给出局部支持区域。证据一致且充分时,门控模块直接接受判断;其余情况交由决策模型结合原图与两路记录处理,并通过强化学习优化决策过程。模型原生输出边界框,评测时由固定的通用分割模型SAM 2(Segment anything model 2)将框转换为掩膜,再计算像素级指标。平台、生成器与处理流程可能同时变化,分布外评测因而不能单独说明某一种分布变化的影响。

利用伪影描述训练解释生成的方法,不依赖前置掩膜作为语言输入。FakeVLM^[40]不设置分割头

与分类头,将伪影识别及其解释统一为图文生成任务。配套数据集 FakeClue 先由 Qwen2-VL、InternVL 与 Deepseek 生成候选伪影描述,再由 Qwen2-VL 聚合为训练文本;解释评测同时使用自动标注与人工标注的测试集。直接生成的文本也可能缺少可见证据,解释的可靠性仍取决于文本与图像证据的对应关系。

语言先验还可直接用于增强定位。Guo 等^[107]提出的 HiFi-Net++ 以预定义伪造类别及文本模板形成输入,通过 CLIP 文本与图像编码器建立空间关联。细化模块调整文本嵌入,并生成各级类别的空间得分图,为定位提供辅助信息。模型输出掩膜与分层伪造属性,不生成自由形式的解释文本,其类别范围仍由预定义分类树限定。

3.3.2 统一框架与视觉中心推理

多项输出的联合建模扩大了取证系统的功能范围,模型在不同数据域中的适用性仍受训练与评测条件影响。ForensicHub 的跨四域评测中,全合成图像检测方法 UnivFD 在图像篡改域的 IMD2020 上取得的图像级 AUC 高于该域的 4 个专用方法^[10]。该基准将掩膜经最大池化转换为图像级标签,只用标签监督,并将图像统一缩放至 256 像素 $\times$ 256 像素,专用方法因而无法沿用原有的像素监督与输入分辨率。跨域比较中,“几乎全部”专用方法表现较低;在图像篡改域内,参与比较的 4 个专用方法均低于 UnivFD。

跨域取证系统可以调用不同领域的专家模型,组合已有检测器的能力。UniShield^[41]设计了感知智能体与检测智能体的两级流程,将输入分派给全合成、Deepfake、图像篡改与文档取证领域的专家模型,再汇总结论并生成统一格式的報告。AgentFoX^[115]先校准各专家的可靠性,并整理其在不同图像类别上的表现;推理时结合这些信息融合存在冲突的判断,输出图像级结论与取证报告。专家协作的效果取决于路由、校准与融合策略,推理成本随实际调用次数变化。对于 UniShield 与 AgentFoX 这类专家协作系统,任务覆盖范围依赖所调用的专家;共享表征及向未见领域的迁移能力仍需相应实验验证。ManipShield^[53]利用大模型内部表征联合处理真伪判断、框定位与解释生成。该方法先通过对比学习微调视觉编码器,再选择真假特征分离较好的语言模型层。该层隐状态分别输入检测、判据属性与区域定位解码器,得到真伪标签、光照或纹理等判据类别以及边界框。3 项结果组织为结构化提示,与图像一同输入大模型生成文字说明。边界框圈定待检区域,不能替代逐像素的篡改范围。

通用大模型在取证中还面临输入处理与训练目标带来的限制。主流视觉编码器以图文对比对齐为训练目标^[98]。部分模型将输入缩放至 336 像素量级,再划分为边长 14 的图像块,重采样可能削弱 JPEG 的 8 $\times$ 8 块栅格与插值周期峰。图文对齐训练更强调语义相关信息,也可能降低特征对细微取证痕迹的敏感度。专用定位器通常保留较高的输入分辨率,例如 IML-ViT 采用 1 024 像素 $\times$ 1 024 像素的原生输入,TruFor 采用 512。Deepfake 基准评测发现,十余个通用与推理型大模型的零样本性能远低于专用检测器,推理链常与真实线索无关^[116]。

外部工具与专用检测器能够为大模型提供额外的取证特征。ForgeryVCR^[42]采用以外部工具为中介的视觉中心推理,由大模型按需调用噪声与频域算子,并将算子输出作为视觉证据输入。模型预测篡改边界框,再由通用分割模型 SAM 2 细化为像素掩膜;主配置使用视觉中间结果与工具调用,不生成中间文字判据,带文字推理的配置用于检测与定位性能的对照。Guo 等^[117]提出的 M2F2-Det 通过桥接模块,将专用人脸检测器的判别特征映射到语言模型的词元空间,使大模型在给出伪造概率的同时生成文字说明。其伪造意图用于特征加权与可视化,不等同于经像素监督评测的篡改掩膜。Kang 等^[43]提出的 LEGION 使用 CLIP 全局图像编码器为真伪分类与解释生成提供特征,另以独立的 SAM 编码器提取定位特征。语言模型在伪影描述后输出专用分割词元,经投影后引导像素解码器生成伪影区域掩

膜。训练时先联合优化定位与解释,再训练分类模块。解释提示涵盖物理、结构与失真3类伪影,分类模块则只判断真伪。这些掩膜刻画全合成图像中的局部伪影,模型还将掩膜与文字说明反馈给生成流程,用于图像重绘与质量改进。

区域信息也可以直接用于约束解释生成。Rethinking VLMs^[118]将定位掩膜同时作为输出目标和辅助输入,以低层定位信号约束语义解释。后训练对解释忠实性的影响取决于监督目标,指令遵循或文本流畅度的改善并不直接证明判据正确;2026年面向AI编辑的推理评测仍发现相关缺口^[119]。外部工具、专用检测器与区域监督提供了不同形式的取证证据,系统性能因而同时受到视觉表征、工具选择与输出模块的影响,各模块的错误也可能传播至后续处理过程。

3.3.3 可解释性的评测现状与不足

解释质量的评测既涉及文本内容,也涉及文本与图像证据的对应关系。Li等^[44]构建的FakeBench按检测、推理与解释3个层次设置问题,除真伪判断外,还要求识别伪影类别并描述其成因,以分别评价判定与解释能力。Ye等^[45]提出的LOKI覆盖图像、视频、三维模型、音频与文本5种模态,在粗粒度判定之外设置细粒度的伪影选择题。这两项工作通过分层问题评测检测与解释能力,但题面及参考答案由人工与闭源模型协同产生,所覆盖的伪影类型仍受标注者先验的限制。

现有解释评测在参考文本来源与评价目标上存在差异。MMTD-Set采用GPT-4o辅助标注,Fake-Clue则聚合多个开放模型的描述。ForgerySleuth^[112]使用经专家修订和交叉核验的ForgeryAnalysis-Eval,评测文本语义相似度,并由GPT-4对推理评分。参考文本的生成及核验过程不同,潜在偏差也有所差异。ForgeryGPT^[39]采用文本一致性指标及人工评价,后者比较阅读解释前后的真伪判断与信心变化。文本与参考答案接近或能够改变使用者判断,均不能单独证明解释来自正确的图像证据;生成与评分模型相同或相近时,还可能引入共同偏差。擦除被指认区域、替换局部内容等扰动检验可以考察判定与解释是否随证据变化,人工核验则有助于识别错误传播。这是对评测方法学不足的指认,并不构成对具体工作的质疑。

表6汇总了各方法的输出与解释质量评测方式。不同方法对自然语言解释、像素掩膜与来源归因的支持存在差异:FakeVLM不输出掩膜,ManipShield提供框定位,HiFi-Net++使用语言先验但不生成文字解释,专家协作方法的定位能力取决于所调用模块。来源标签也可能只覆盖预定义类别。分割词元引导解码的机制可追溯至Lai等^[120]提出的LISA,其将语言模型中的分割词元映射为掩膜解码器的提示。掩膜质量与推理成本仍受架构及实验条件影响,采用大模型并不保证两者同时改善;Forensic-Zip的词元效率研究进一步考察了取证表征与计算开销之间的取舍^[121]。

3.4 判别依据与适用条件的综合比较

3条研究路线在任务目标与架构上存在交叉,其差异主要体现在判别依据与证据组织方式上。噪声域与相机指纹、频域与压缩痕迹、边界与语义不一致主要描述图像中可观测的差异;重建与扩散反演残差依赖输入与外部生成模型的关系;预训练语义特征提供特征空间中的判别依据;视觉证据与语言推理则组织来自图像或工具的证据并生成解释。语言推理用于组织证据,本身不构成独立的物理痕迹。

3类伪造源对这些依据的影响如表7所示。传统局部篡改可能保留区域间的成像差异;生成式局部编辑还受到背景保留、区域融合和编辑范围的影响;全合成图像缺少原始拍摄背景,判别更多依赖生成伪影、特征分布或重建差异。同一方法可以结合多种依据,其效果取决于这些依据在具体图像中的有效性及相互补充程度。

表7 不同判别依据与证据组织方式的适用条件

Table 7 Applicability of different discriminative cues and evidence organization strategies

判别依据或证据组织方式	传统局部篡改	生成式局部编辑	全合成图像
噪声域与相机指纹	区域残差可分	需保留背景残差参照	无原始相机指纹
频域与压缩痕迹	量化或频谱存在差异	取决于融合与压缩	可分析生成频谱伪影
边界与语义不一致	边界或内容存在异常	取决于融合与编辑内容	可分析全局内容异常
重建与扩散反演残差	需验证残差可分性	需局部残差可分	需验证反演器匹配
预训练语义特征	对轻微局部编辑受限	需识别局部特征偏离	需跨生成器特征可分
视觉证据与语言推理	依赖视觉或工具证据	定位依赖具体模块	侧重判定与证据解释

注:按主要判别依据归纳,同一方法可涉及多行,不表示统一性能排序。全合成图像的真值掩膜为全1,不区分图内编辑区域。

4 评测协议与方法性能的可比性

跨文献的性能比较需要考虑数据集构成、指标计算和阈值选取的一致性。同一方法在不同设置下可能获得不同结果,评测差异也可能改变方法排名,但不能据此认定原报告数字普遍虚高。第4.1节介绍主要评测差异及其对结果的影响,第4.2节归纳可比评测所需的信息,并讨论计算成本与性能比较的适用条件。

4.1 评测设置的分歧与排名的可逆转性

针对数据集构成、指标定义与阈值选取在各文献中不统一的问题,统一基准相继出现,覆盖范围由单一取证域扩展到多域:DeepfakeBench^[122]整合9个人脸伪造数据集与15个方法,IMDL-BenCo^[55]复现8个方法与6个主干、采用固定阈值0.5并配两套训练协议,ForensicHub^[10]跨4类取证任务、23个数据集与42个模型。其中IMDL-BenCo的Protocol-MVSS仅用CASIA v2训练,在5个数据集上测试;Protocol-CAT则在CASIA v2、Fantastic Reality、IMD2020与两个自动合成库的混合集上训练,并从测试集中排除已用于训练的IMD2020,因此只在其余4个数据集上测试。

在数据集构成方面,表2列出了文献间的不同记载。NIST16的真图数分别被记为0与560,部分工作据前一种记载将其排除在图像级评测之外;CASIA v1也被以缺少真图为由排除。CASIA v2的真图与伪图数有7 200与5 123、7 491与5 123、7 491与5 063三种记载,总样本数有12 323、12 554与12 641三个数值,后两项分别见于同一团队的两篇基准工作^[10,55]。这些差异可能与发布版本或评测子集有关,现有信息尚不能确定具体原因。

指标计算方式也会影响结果解释。IMDL-BenCo^[55]归纳了两个变体:invert- F_1 将预测掩膜整体取反后再与真值计算 F_1 ,permute- F_1 则取原始值与取反值中的较大者,相当于在评测时从预测掩膜及其补集中选择更优结果。当预测掩膜恰为真值补集时,原始 F_1 为0.00,但取反后与真值重合,permute- F_1 仍可获得高分。不同指标的数值尺度同样存在差异。例如,ManTra-Net在Protocol-MVSS下的平均像素级AUC为0.783,同批测试集上的平均交并比为0.141,两者相差5.5倍,但这一倍数不表示定位能力的同等差距。评测范围也需加以区分:如2.2节所述,仅在伪造图像上计算的像素级 F_1 能反映篡改图内的背景误报,却不计完整真实图像上的假阳性。指标变体、数值尺度与样本范围共同影响结果的可比性。

重训练实验反映了原报告与复现结果之间的差异。IML-ViT的作者^[25]同时列出MVSS-Net的原文报告值与重训值,平均像素级 F_1 由0.394降至0.270,降幅为31.5%;其中Columbia上由0.638降至0.303,降幅为52.5%。在IMDL-BenCo的统一复现中,ManTra-Net、MVSS-Net与NCL-IML在Columbia上的降幅为33%~40%,在COVERAGE上为41%~69%^[55]。不过变化方向随数据集而异:

MVSS-Net在CASIAv1上的复现值由0.452升至0.534,比原文报告值高18.1%。因此,复现差异不能概括为原报告普遍虚高。可配对比较的样本也有限,8个被复现方法中只有4个的训练集与指标定义一致,其余方法因训练配方不同或仅报告微调后AUC(如PSCC-Net与ObjectFormer)而难以直接比较。IMDL-BenCo还发布了按结构相似度去重的版本,以减少随机划分及微调可能造成的数据泄漏,更换版本后各方法性能有升有降。

评测条件的变化对不同方法并非产生相同影响,因此还可能改变相对排名。在IMDL-BenCo的Protocol-MVSS下,TruFor的平均像素级AUC与交并比分别为0.889与0.478,均高于IML-ViT;在Protocol-CAT下,IML-ViT则以0.942与0.678高于TruFor。两套协议同时改变了训练数据和测试集构成,原文报告的五库均值与四库均值之差不能单独归因于训练配方。

阈值选取也会影响排名。MVSS-Net原文^[5]同时报告了最优阈值和固定0.5阈值的结果:ManTra-Net由0.649(第2)降至0.192(第5),GSR-Net由0.504(第5)变为0.324(第2),两者名次互换。TruFor原文^[7]给出14个方法在8个数据集上的配对结果,按跨数据集平均值排序时,10个方法的名次发生变化。单个数据集上的最大降幅出现在PSCC-Net,其在OpenForensics^[123]上由0.353降至0.105,降幅为70.3%。固定阈值的适用性还与输出含义有关:TruFor原文说明,Splicebuster与Noiseprint输出任意值域的异常图,评测前需要归一化,固定0.5阈值不具有概率判定的含义。不同训练与测试设置、阈值选取方式下的排名变化如图8所示。

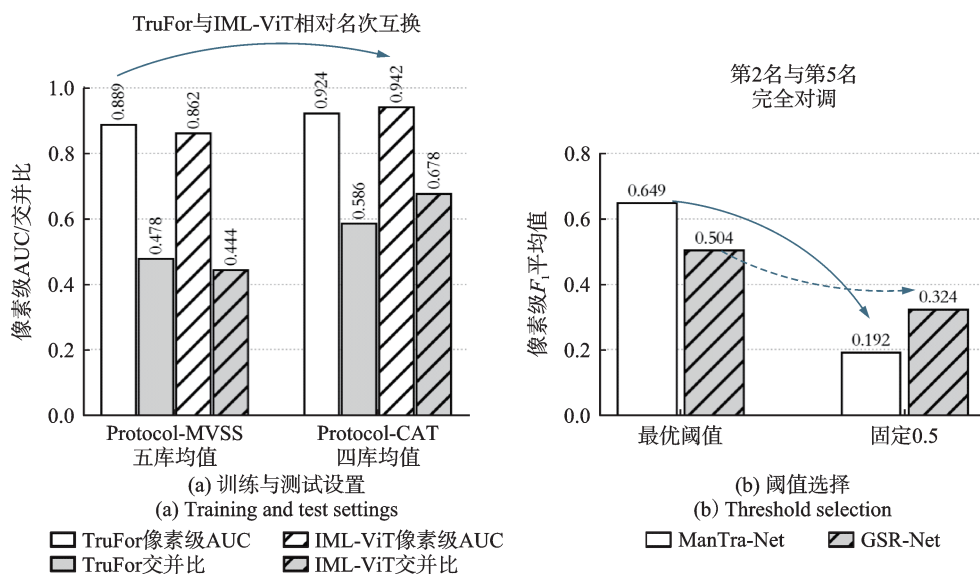


图8 不同评测设置下的方法排名变化

Fig.8 Changes in method rankings under different evaluation settings

4.2 面向可比性与可部署性的评测建议

结果的可比性取决于实验设置是否一致,可比评测至少涉及5项信息:(1)数据集版本与真伪构成,包括真图数量及其是否参与评测;(2)指标变体,包括 F_1 的平均方式、是否取反或取较大值,以及AUC的像素级或图像级计算范围;(3)阈值选取方式,区分固定0.5、逐图最优与逐测试集最优,并说明异常图的归一化方式;(4)训练配方,包括预训练数据、训练集构成与输入分辨率;(5)退化条件,包括压缩与缩放等后处理。

只在伪造图像上计算的像素级指标不能反映完整真实图像上的误报,结合含真图负样本的图像级指标,可以补充衡量真实图像被错判为伪造的情况。部署还需考虑计算成本。ForensicHub中,ConvNeXt以50.22M参数、22.74G浮点运算量取得0.822的平均图像级AUC;MVSS-Net的参数量与运算量分别为147.00M、83.08G,结果为0.736^[10]。后者的参数量与运算量约为前者的3倍,但该评测下的图像级性能较低。此比较受到实验条件限制:基准通过最大池化将掩膜转换为图像级标签,只用标签监督,并统一缩放到256像素 $\times$ 256像素;IML-ViT的原生输入为1 024、TruFor为512,专用方法未能沿用其像素监督与分辨率设置。因此,这一结果仅说明该协议下的计算成本与图像级性能关系,不能直接推广到像素级定位或不同分辨率的应用。边缘侧初筛与平台侧复核需要结合任务、分辨率和监督条件选择方法,退化条件也应与实际传输过程相符。

5 传输链路退化下的鲁棒性

真实场景中,图像从生成或编辑到被检测,可能经历压缩、缩放与多次转发。即使伪造区域保持不变,承载判别信息的像素统计也可能改变,使检测器产生误报或漏检。退化增广、平台处理仿真与一致性约束分别从训练样本、传播过程和预测关系入手,减小处理造成的预测变化,但不能恢复已被不可逆处理完全删除的取证信息。

5.1 传输退化与取证痕迹的变化

压缩、缩放和噪声扰动作用于不同层次的取证信息。JPEG再压缩改变频域系数及量化历史,可能削弱已有压缩差异;缩放通过插值与重采样改变局部邻域关系,较小的编辑区域及边界也可能被平滑;附加噪声则可能掩盖相机残差或产生额外的异常响应。相关退化会同时作用于真实区和伪造区,因此检测器面对的不仅是篡改痕迹减弱,还包括真实背景出现与训练时不同的统计特征^[7,87]。

面向社交网络的研究以可预测噪声与未见噪声描述平台传输的不确定性^[124-125]。前者对应可通过已知操作或平台输入输出样本近似建模的变化;后者描述仿真未能覆盖且可能影响检测器的扰动,不等同于某一种固定的噪声。连续转发还会叠加多种操作,处理顺序与参数会共同决定最终图像。因此,仅考察模型对单一质量因子或固定缩放倍率的适应性,难以全面反映其对实际复杂传输链路的鲁棒性。

5.2 面向传输退化的鲁棒训练

退化增广在训练中纳入经压缩、缩放等操作处理的样本,使检测器同时学习原始和退化图像。TruFor^[7]在Noiseprint++的预训练中使用经过多种不同处理的图像,使残差表征对相机内处理和外部编辑产生响应,并在定位网络训练中采用压缩与缩放增广。同一内容的不同处理版本有助于模型学习较稳定的判别特征。增广能扩大已知操作的覆盖范围,但不能保证对训练范围外的操作组合仍然有效。

平台仿真利用传输前后的配对图像学习退化过程。Wu等^[124]通过包含残差学习与可微JPEG层的网络近似平台处理,再结合增大检测误差的对抗扰动训练定位器。模拟器生成接近已观察平台的样本,对抗扰动补充较难建模的情况;相关研究还在实际平台传输图像上验证了检测性能^[125]。对抗扰动用于近似训练中的未知变化,其分布未必与真实平台处理一致,跨平台及平台更新后的有效性仍需相应测试验证。

一致性训练则约束不同退化版本的预测与表征。ProCoS^[87]以未施加额外退化的图像作为教师分支输入,将随机退化后的图像输入学生分支,通过一致性约束减小定位结果的漂移;教师参数由学生参数的指数移动平均更新。该方法同时维护篡改与背景的特征原型,使同类像素更紧凑、不同类别更易区分。其效果仍受场景、退化组合与教师预测质量影响。

5.3 公开实验中的退化表现与适用范围

不同取证依据对传输处理的敏感性有所区别,与表7归纳的适用条件相对应。PIM的跨库直测^[34]中,AutoSplice质量因子由100降至75时,学习压缩伪影的CAT-Net像素级AUC从0.887降至0.607,降幅约31.6%,为该实验17个方法中的最大降幅。相同压缩梯度下,各方法的像素级AUC平均下降约12%,MCC下降约50%;MVSS-Net++的AUC下降14.6%,MCC下降80.7%。AUC与MCC的变化幅度不同,仅使用AUC可能低估固定阈值下预测结果的退化。表8列出其中代表方法在各压缩档的AUC与MCC。

表8 JPEG压缩下代表方法的像素级定位性能

Table 8 Pixel-level localization performance of representative methods under JPEG compression

方法	AutoSplice 质量因子 100		AutoSplice 质量因子 90		AutoSplice 质量因子 75	
	AUC	MCC	AUC	MCC	AUC	MCC
ManTra-Net ^[23]	0.664	0.189	0.626	0.160	0.660	0.177
SPAN ^[77]	0.645	0.020	0.549	0.005	0.566	0.007
PSCC-Net ^[56]	0.749	0.275	0.657	0.195	0.630	0.156
MVSS-Net++ ^[76]	0.836	0.280	0.751	0.101	0.714	0.054
CAT-Net ^[6]	0.887	0.578	0.720	0.301	0.607	0.163
TruFor ^[7]	0.827	0.382	0.818	0.358	0.820	0.367
PIM ^[34]	0.940	0.715	0.852	0.402	0.729	0.151

注:数据取自PIM跨库评测;AutoSplice未参与训练,MCC阈值为0.5。原测17个方法,选列7个。

模型间的差异还与预训练和架构有关。从质量因子100到75,CAT-Net与TruFor的AUC降幅分别为31.6%和0.8%;TruFor的抗压缩表现部分归因于Noiseprint++的额外大规模预训练,不能仅用噪声域特征解释^[34]。这组比较涉及的17个方法使用不同训练数据与架构,因而不能将性能差异单独归因于取证特征,也不能排除模型容量的影响。含DCT分支的Mesorch在中等压缩下仍有较好表现,部分条件下性能有所提高,在低质量档才明显退化^[87],说明频域特征的性能变化未必单调,也并非始终比其他特征更易受损。

噪声与复合退化实验补充了压缩测试的结果。在ProCoS的单一退化实验中^[87],采用Noiseprint++噪声残差的TruFor在高斯噪声下表现出同批方法中最明显的像素级 F_1 与交并比下降,噪声可能掩盖残差中的判别信息,训练分布与架构差异也可能影响这一结果。该研究还将高斯噪声、高斯模糊与JPEG压缩3种算子组合为9种复合退化,ProCoS的平均像素级 F_1 衰减为0.0873,最大衰减为0.1966。这些 F_1 衰减主要来自自建的特种设备检验图像,反映的是这一特定场景下的性能,不能视为通用评测结果。

6 挑战与展望

面向实际应用,检测与定位仍需解决跨域泛化、来源归因、解释评测及链路鲁棒性等问题。以下从6个方面说明现有方法的局限与可继续验证的方向。

(1)跨域表征与统一评测。ForensicHub的统一跨域评测显示,不同任务之间的性能迁移仍不稳定。在图像篡改数据集上,采用冻结CLIP特征的全合成检测器优于多个专用方法^[10],但统一缩放与监督设置限制了专用方法原有的条件,因此结果不能推广为专用方法普遍无效。不同领域的有效特征能否共

享, 仍需在可比的分辨率与监督条件下验证。未来可结合4类取证任务开展共享表征研究, 分析跨域训练带来的特征互补及负迁移。

(2) 未见生成器的开集泛化与持续学习。HiFi-Net通过4层分类树逐级预测来源; ForensicHub另报告了其跨域图像级检测结果, 平均AUC低于同批多个方法^[10]。对于训练类别之外的新生成器, 封闭集分类器缺少对应标签, 仍会将其分配到已有类别。与此同时, 上采样算子与解码器等实现细节的变化可能改变模型指纹, 使来源归因同时面临特征分布变化与类别扩展。未来需要结合开集拒识与增量学习, 研究在生成器更新后仍具有区分能力的归因特征。

(3) 可解释性的可信度量。FakeShield的MMTD-Set采用GPT-4o辅助标注^[9], FakeVLM的FakeClue则通过多个开放模型生成并聚合描述^[40]。参考文本的来源不同, 潜在偏差也有所差异。未来应在文本相似度与评分模型评价之外, 增加人工证据核验、局部擦除与反事实编辑检验, 考察解释是否随相关图像证据变化; 使用同类模型生成与评分时, 需报告并控制共同偏差。

(4) 评测设置的标准化。训练数据、指标变体与阈值选取均可能改变方法排名。MVSS-Net原文将逐数据集最优阈值改为固定阈值后, 原排名第2与第5的两个方法发生名次互换^[5]。permute- F_1 在原始掩膜及其补集中选择较优结果, 也改变了指标的含义。在Protocol-MVSS与Protocol-CAT下, TruFor和IML-ViT的相对名次也发生变化, 但这两套协议同时改变了训练数据与测试集构成^[55]。这些变化随数据集和方法而异, 难以在实验结束后统一校正。未来需要明确数据集版本与真伪构成、指标变体、阈值选取及训练数据等基本信息, 形成可复用的评测协议。

(5) 被动取证与主动溯源的协同。主动溯源包括容器级凭证、权重级水印与采样级水印等方式, C2PA内容凭证已迭代至2.4版^[36]。这些技术依赖生成或发布方配合, 覆盖范围因而有限。截图与平台后处理可能移除容器级清单, 替换回原始解码器也可能使权重级水印失效, 凭证缺失本身不能作为真伪依据。未来可将来源凭证校验与被动取证结合, 对缺少有效凭证的图像继续进行检测与定位, 并比较不同验证方式的覆盖范围与计算成本。

(6) 反取证对抗与真实链路鲁棒性。真实分发可能叠加平台压缩、缩放与多次转发, 其处理组合和参数还会随平台更新而变化。固定退化条件下的结果^[34]尚不足以说明模型对这些变化的适应能力。多模态引导的通用反取证扰动可以针对多类取证器构造^[126], 再压缩与再生成也可能削弱已有痕迹或水印^[110]。未来需要在实际平台和复合操作下验证鲁棒训练的有效性, 区分自然传输退化与主动攻击, 并报告攻击知识、预算及正常图像误报情况。

7 结束语

本文从数据与方法两个方面梳理了图像伪造检测与定位的研究进展, 整理3类伪造源的痕迹、常用数据集与评价指标, 并沿传统取证线索、生成式伪造与多模态大模型3条研究路线说明代表方法的机制与适用条件。在数据与评测方面, 常用数据集的规模与标注粒度存在差异, 部分数据集的真伪构成在各文献中记载不一。训练数据、监督方式与阈值选取均会影响方法排名, 因此跨论文比较需要明确实验条件, 并分别评价图像级判定与像素级定位能力。在判别依据方面, 经典噪声、压缩与边界线索的有效性取决于编辑后背景、区域差异及处理痕迹的保留程度。生成器伪影、扩散反演残差与预训练特征提供了补充依据, 但其泛化能力仍受到目标数据分布的影响。现有实验显示, 不同方法在退化条件下的性能变化与所用线索有关, 训练数据及架构差异的作用尚未充分区分, 因此单项实验结果仍需结合具体条件解释。在输出能力方面, 多模态大模型支持检测、定位、归因与文字解释的多种组合, 各方法对这些任务的覆盖程度不同。注意图、定位掩膜与文字说明具有不同含义, 不能相互替代。全合成图

像的真值掩膜为全1,不能区分图内编辑区域,评测重点相应转向判定与归因。自然语言解释则需要分别评价其可读性及对图像证据的忠实程度。

后续研究可重点检验跨生成器与跨编辑流程的泛化能力,完善能够区分输出任务的评测协议,并在真实复合退化下比较特征互补与计算开销。将方法的适用条件、输出能力和部署成本一并报告,有助于根据具体应用选择检测与定位方案。

参考文献:

- [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] FRIDRICH J, KODOVSKY J. Rich models for steganalysis of digital images[J]. *IEEE Transactions on Information Forensics and Security*, 2012, 7(3): 868-882.
- [3] COZZOLINO D, POGGI G, VERDOLIVA L. Splicebuster: A new blind image splicing detector[C]//*Proceedings of 2015 IEEE International Workshop on Information Forensics and Security (WIFS)*. Rome, Italy: IEEE, 2015: 1-6.
- [4] ZHOU P, HAN X, MORARIU V I, et al. Learning rich features for image manipulation detection[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA: IEEE, 2018: 1053-1061.
- [5] CHEN X, DONG C, JI J, et al. Image manipulation detection by multi-view multi-scale supervision[C]//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE, 2021: 14165-14173.
- [6] 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.
- [7] GUILLARO F, COZZOLINO D, SUD A, et al. TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization[C]//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC, Canada: IEEE, 2023: 20606-20615.
- [8] GUO X, LIU X, REN Z, et al. Hierarchical fine-grained image forgery detection and localization[C]//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC, Canada: IEEE, 2023: 3155-3165.
- [9] XU Z, ZHANG X, LI R, et al. FakeShield: Explainable image forgery detection and localization via multi-modal large language models[C]//*Proceedings of the International Conference on Learning Representations (ICLR 2025)*. [S.l.]: [s.n.], 2025: 31186-31216.
- [10] DU B, ZHU X, MA X, et al. ForensicHub: A unified benchmark & codebase for all-domain fake image detection and localization[C]//*Proceedings of Advances in Neural Information Processing Systems 38*. San Diego, California, USA and Mexico City, Mexico: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2025: 72073-72102.
- [11] VERDOLIVA L. Media forensics and DeepFakes: An overview[J]. *IEEE Journal of Selected Topics in Signal Processing*, 2020, 14(5): 910-932.
- [12] KADHA V, BAKSHI S, DAS S K. Unravelling digital forgeries: A systematic survey on image manipulation detection and localization[J]. *ACM Computing Surveys*, 2025, 57(12): 1-36.
- [13] 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.
- [14] 李美玲, 钱振兴, 张新鹏. 人工智能生成图像检测技术综述[J]. *中国图象图形学报*, 2026, 31(1): 13-44.
LI Meiling, QIAN Zhenxing, ZHANG Xinpeng. Survey on artificial intelligence-generated image detection[J]. *Journal of Image and Graphics*, 2026, 31(1): 13-44.
- [15] ZOU Y, LI P, LI Z, et al. Survey on AI-generated media detection: From non-MLLM to MLLM[EB/OL]. (2025-02-12). <https://doi.org/10.48550/arXiv.2502.05240>.

- [16] 彭春蕾, 李俊桦, 刘德成, 等. 大模型时代的深度伪造检测[J]. 中国科学: 信息科学, 2026, 56(1): 1-22.
PENG Chunlei, LI Junye, LIU Decheng, et al. Deepfake detection in the era of large models[J]. Scientia Sinica (Informationis), 2026, 56(1): 1-22.
- [17] 刘晓龙, 刘欢, 赵耀, 等. AIGC 伪造内容被动检测与主动防御技术综述[J]. 中国科学(信息科学), 2025, 55(9): 2250-2288.
LIU Xiaolong, LIU Huan, ZHAO Yao, et al. Review of AIGC passive detection and proactive protection[J]. Scientia Sinica (Informationis), 2025, 55(9): 2250-2288.
- [18] WU X, CHEN X, WU X, et al. Explainable image-centric forgery detection: A survey[EB/OL]. (2025-10-21). <https://doi.org/10.36227/techrxiv.176101439.91738583/v1>.
- [19] COZZOLINO D, VERDOLIVA L. Noiseprint: A CNN-based camera model fingerprint[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 144-159.
- [20] WU Y, ABD-ALMAGEED W, NATARAJAN P. BusterNet: Detecting copy-move image forgery with source/target localization[C]//Proceedings of Computer Vision—ECCV 2018. Cham, Switzerland: Springer International Publishing, 2018: 170-186.
- [21] MAREEN H, VANDEN BUSSCHE D, GUILLARO F, et al. Comprint: Image forgery detection and localization using compression fingerprints[C]//Proceedings of Pattern Recognition, Computer Vision, and Image Processing, ICPR 2022 International Workshops and Challenges. Cham, Switzerland: Springer Nature, 2023: 281-299.
- [22] GUO K, ZHU H, CAO G. Effective image tampering localization via enhanced transformer and co-attention fusion[C]//Proceedings of ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Korea: IEEE, 2024: 4895-4899.
- [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 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 9535-9544.
- [24] WANG J, WU Z, CHEN J, et al. ObjectFormer for image manipulation detection and localization[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 2354-2363.
- [25] MA X, DU B, JIANG Z, et al. IML-ViT: Benchmarking image manipulation localization by vision transformer[EB/OL]. (2024-11-24). <https://doi.org/10.48550/arXiv.2307.14863>.
- [26] SU L, MA X, ZHU X, et al. Can we get rid of handcrafted feature extractors? SparseViT: Nonsemantics-centered, parameter-efficient image manipulation localization through sparse-coding transformer[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, DC, USA: AAAI, 2024: 7024-7032.
- [27] ZHU X, MA X, SU L, et al. Mesoscopic insights: Orchestrating multi-scale hybrid architecture for image manipulation localization[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, DC, USA: AAAI, 2024: 11022-11030.
- [28] WANG S Y, WANG O, ZHANG R, et al. CNN-generated images are surprisingly easy to spot for now[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020: 8692-8701.
- [29] TAN C, LIU H, ZHAO Y, et al. Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection[C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2024: 28130-28139.
- [30] OJHA U, LI Y, LEE Y J. Towards universal fake image detectors that generalize across generative models[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE, 2023: 24480-24489.
- [31] YAN S, LI O, CAI J, et al. A sanity check for AI-generated image detection[C]//Proceedings of the International Conference

- on Learning Representations (ICLR). [S.l.]: [s.n.], 2025: 70702-70720.
- [32] WANG Z, BAO J, ZHOU W, et al. DIRE for diffusion-generated image detection[C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023: 22388-22398.
- [33] CHEN Baoying, ZENG Jishen, YANG Jianquan, et al. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images[C]//Proceedings of the 41st International Conference on Machine Learning. [S.l.]: [s.n.], 2024: 7621-7639.
- [34] KONG C, LUO A, WANG S, et al. Pixel-inconsistency modeling for image manipulation localization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(6): 4455-4472.
- [35] WANG M, NIU S, ZHANG J. ForgDiffuser: General image forgery localization with diffusion models[C]//Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. Montreal, Canada: International Joint Conferences on Artificial Intelligence Organization, 2025: 1954-1962.
- [36] Coalition for Content Provenance and Authenticity. C2PA technical specification: Version 2.4[EB/OL]. [2026-09-05]. <https://c2pa.org/specifications>.
- [37] ZHANG X, LI R, YU J, et al. EditGuard: Versatile image watermarking for tamper localization and copyright protection[C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2024: 11964-11974.
- [38] WEN Y, KIRCHENBAUER J, GEIPING J, et al. Tree-rings watermarks: Invisible fingerprints for diffusion images[C]//Proceedings of Advances in Neural Information Processing Systems 36. New Orleans, Louisiana, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 58047-58063.
- [39] ZHANG F, LIU J, ZHU J, et al. ForgeryGPT: A multimodal LLM for interpretable image forgery detection and localization [EB/OL]. (2026-04-07). <https://doi.org/10.48550/arXiv.2410.10238>.
- [40] WEN S, YE J, FENG P, et al. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation[C]//Proceedings of Advances in Neural Information Processing Systems 38. San Diego, California, USA and Mexico City, Mexico: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2025: 65691-65724.
- [41] HUANG Q, XU Z, ZHANG X, et al. UniShield: An adaptive multi-agent framework for unified forgery image detection and localization[EB/OL]. (2026-05-15). <https://doi.org/10.48550/arXiv.2510.03161>.
- [42] WANG Y, CHEN S, WANG H, et al. ForgeryVCR: Visual-centric reasoning via efficient forensic tools in MLLMs for image forgery detection and localization[EB/OL]. (2026-08-12). <https://doi.org/10.48550/arXiv.2602.14098>.
- [43] KANG H, WEN S, WEN Z, et al. LEGION: Learning to ground and explain for synthetic image detection[C]//Proceedings of 2025 IEEE/CVF International Conference on Computer Vision (ICCV). Honolulu, HI, USA: IEEE, 2025: 18937-18947.
- [44] LI Y, LIU X, WANG X, et al. FakeBench: Probing explainable fake image detection via large multimodal models[J]. IEEE Transactions on Information Forensics and Security, 2025, 20: 8730-8745.
- [45] YE J, ZHOU B, HUANG Z, et al. LOKI: A comprehensive synthetic data detection benchmark using large multimodal models [EB/OL]. (2024-10-13). <https://doi.org/10.48550/arXiv.2410.09732>.
- [46] DONG J, WANG W, TAN T. CASIA image tampering detection evaluation database[C]//Proceedings of 2013 IEEE China Summit and International Conference on Signal and Information Processing. Beijing, China: IEEE, 2013: 422-426.
- [47] HSU Y F, CHANG S F. Detecting image splicing using geometry invariants and camera characteristics consistency[C]//Proceedings of 2006 IEEE International Conference on Multimedia and Expo. Toronto, ON, Canada: IEEE, 2006: 549-552.
- [48] GUAN H, KOZAK M, ROBERTSON E, et al. MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation[C]//Proceedings of 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). Waikoloa Village, HI, USA: IEEE, 2019: 63-72.
- [49] NOVOZAMSKY A, MAHDIAN B, SAIC S. IMD2020: A large-scale annotated dataset tailored for detecting manipulated images[C]//Proceedings of 2020 IEEE Winter Applications of Computer Vision Workshops (WACVW). Snowmass Village,

CO, USA: IEEE, 2020: 71-80.

- [50] MAHFOUDI G, TAJINI B, RETRAINT F, et al. DEFACTO: Image and face manipulation dataset[C]//Proceedings of 2019 27th European Signal Processing Conference (EUSIPCO). A Coruna, Spain: IEEE, 2019: 1-5.
- [51] JIA S, HUANG M, ZHOU Z, et al. AutoSplice: A text-prompt manipulated image dataset for media forensics[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Vancouver, BC, Canada: IEEE, 2023: 893-903.
- [52] CHEN Y, HUANG X, ZHANG Q, et al. GIM: A million-scale benchmark for generative image manipulation detection and localization[C]//Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence. Washington, DC, USA: AAAI, 2025, 39(2): 2311-2319.
- [53] XU Z, DUAN H, WANG X, et al. ManipShield: A unified framework for image manipulation detection, localization and explanation[EB/OL]. (2026-01-19). <https://doi.org/10.48550/arXiv.2511.14259>.
- [54] ZHU M, CHEN H, YAN Q, et al. GenImage: A million-scale benchmark for detecting AI-generated image[C]//Proceedings of Advances in Neural Information Processing Systems 36. New Orleans, Louisiana, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 77771-77782.
- [55] MA X, ZHU X, SU L, et al. IMDL-BenCo: A comprehensive benchmark and codebase for image manipulation detection & localization[C]//Proceedings of Advances in Neural Information Processing Systems 37. Vancouver, BC, Canada: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024: 134591-134613.
- [56] LIU X, LIU Y, 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.
- [57] WEN B, ZHU Y, SUBRAMANIAN R, et al. COVERAGE: A novel database for copy-move forgery detection[C]//Proceedings of 2016 IEEE International Conference on Image Processing (ICIP). Phoenix, AZ, USA: IEEE, 2016: 161-165.
- [58] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 10674-10685.
- [59] NICHOL A Q, DHARIWAL P, RAMESH A, et al. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models[C]//Proceedings of the International Conference on Machine Learning. [S.l.]: [s.n.], 2022: 16784-16804.
- [60] WANG Y, YU J, ZHANG J. Zero-shot image restoration using denoising diffusion null-space model[EB/OL].(2022-12-01). <https://doi.org/10.48550/arXiv.2212.00490>.
- [61] KARRAS T, AILA T, LAINE S, et al. Progressive growing of GANs for improved quality, stability, and variation[EB/OL]. (2017-10-27). <https://doi.org/10.48550/arXiv.1710.10196>.
- [62] 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.
- [63] 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.
- [64] BAMMEY Q. Synthbuster: Towards detection of diffusion model generated images[J]. *IEEE Open Journal of Signal Processing*, 2024, 5: 1-9.
- [65] TRALIC D, ZUPANCIC I, GRGIC S, et al. CoMoFoD: New database for copy-move forgery detection[C]//Proceedings ELMAR-2013. Zadar, Croatia: IEEE, 2013: 49-54.
- [66] 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.
- [67] ZAMPOGLOU M, PAPADOPOULOS S, KOMPATSIARIS Y. Detecting image splicing in the wild (WEB)[C]//Proceedings of 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). Turin, Italy: IEEE, 2015:

1-6.

- [68] HELLER S, ROSSETTO L, SCHULDT H. The PS-battles dataset—An image collection for image manipulation detection [EB/OL]. (2018-04-13). <https://doi.org/10.48550/arXiv.1804.04866>.
- [69] KNIAZ V V, KNYAZ V A, REMONDINO F. The point where reality meets fantasy: Mixed adversarial generators for image splice detection[C]//Proceedings of Advances in Neural Information Processing Systems. [S.l.]: Curran Associates, Inc., 2019, 32: 215-226.
- [70] KORUS P, HUANG J. Multi-scale analysis strategies in PRNU-based tampering localization[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(4): 809-824.
- [71] ZHOU P, CHEN B C, HAN X, et al. Generate, segment, and refine: Towards generic manipulation segmentation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2020: 12811-12818.
- [72] BAYAR B, STAMM M C. A deep learning approach to universal image manipulation detection using a new convolutional layer [C]//Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security. Vigo Galicia, Spain: ACM, 2016: 5-10.
- [73] HUH M, LIU A, OWENS A, et al. Fighting fake news: Image splice detection via learned self-consistency[C]//Proceedings of Computer Vision—ECCV 2018. Cham, Switzerland: Springer International Publishing, 2018: 106-124.
- [74] MAYER O, STAMM M C. Forensic similarity for digital images[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 1331-1346.
- [75] WANG J, SUN K, CHENG T, et al. Deep high-resolution representation learning for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3349-3364.
- [76] DONG C, CHEN X, HU R, 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.
- [77] HU X, ZHANG Z, JIANG Z, et al. SPAN: Spatial pyramid attention network for image manipulation localization[C]//Proceedings of Computer Vision—ECCV 2020. Cham, Switzerland: Springer International Publishing, 2020: 312-328.
- [78] NILOY F F, KUMAR BHAUMIK K, WOO S S. CFL-Net: Image forgery localization using contrastive learning[C]//Proceedings of 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, HI, USA: IEEE, 2023: 4631-4640.
- [79] HAO J, ZHANG Z, YANG S, et al. TransForensics: Image forgery localization with dense self-attention[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021: 15035-15044.
- [80] XIE E, WANG W, YU Z, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers[C]//Proceedings of Advances in Neural Information Processing Systems. [S.l.]: [s.n.], 2021, 34: 12077-12090.
- [81] HE K, CHEN X, XIE S, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 15979-15988.
- [82] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[EB/OL]. (2020-10-22). <https://doi.org/10.48550/arXiv.2010.11929>.
- [83] LIU Z, MAO H, WU C Y, et al. A ConvNet for the 2020s[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 11966-11976.
- [84] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021: 9992-10002.
- [85] TAN M, LE Q V. EfficientNet: Rethinking model scaling for convolutional neural networks[C]//Proceedings of the International Conference on Machine Learning. [S.l.]: [s.n.], 2019: 6105-6114.
- [86] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 770-778.
- [87] CHEN G, LIU G, LUO W. ProCoS: Degradation-robust manipulation localization for special equipment inspection images[J].

Sensors, 2026, 26(12): 3647.

- [88] BI X, WEI Y, XIAO B, et al. RRU-Net: The ringed residual U-Net for image splicing forgery detection[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, CA, USA: IEEE, 2019: 30-39.
- [89] YANG C, LI H, LIN F, et al. Constrained R-CNN: A general image manipulation detection model[C]//Proceedings of 2020 IEEE International Conference on Multimedia and Expo (ICME). London, United Kingdom: IEEE, 2020: 1-6.
- [90] ZHOU J, MA X, DU X, et al. Pre-training-free image manipulation localization through non-mutually exclusive contrastive learning[C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023: 22289-22299.
- [91] DURALL R, KEUPER M, KEUPER J. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020: 7887-7896.
- [92] FRANK J, EISENHOFER T, SCHÖNHERR L, et al. Leveraging frequency analysis for deep fake image recognition[C]//Proceedings of the 37th International Conference on Machine Learning. [S.l.]: ACM, 2020: 3247-3258.
- [93] TAN C, TAO R, LIU H, et al. C2P-CLIP: Injecting category common prompt in CLIP to enhance generalization in deepfake detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2025: 7184-7192.
- [94] TAN C, ZHAO Y, WEI S, et al. Learning on gradients: Generalized artifacts representation for GAN-generated images detection[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE, 2023: 12105-12114.
- [95] TAN C, ZHAO Y, WEI S, et al. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, DC, USA: AAAI, 2024: 5052-5060.
- [96] CORVI R, COZZOLINO D, ZINGARINI G, et al. On the detection of synthetic images generated by diffusion models[C]//Proceedings of ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023: 1-5.
- [97] ZHONG N, XU Y, LI S, et al. PatchCraft: Exploring texture patch for efficient AI-generated image detection[EB/OL]. (2024-03-07). <https://doi.org/10.48550/arXiv.2311.12397>.
- [98] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the International Conference on Machine Learning. [S.l.]: [s.n.], 2021: 8748-8763.
- [99] LIU H, TAN Z, TAN C, et al. Forgery-aware adaptive transformer for generalizable synthetic image detection[C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2024: 10770-10780.
- [100] YU N, DAVIS L, FRITZ M. Attributing fake images to GANs: Learning and analyzing GAN fingerprints[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019: 7555-7565.
- [101] DHARIWAL P, NICHOL A. Advances in neural information processing systems[C]//Proceedings of Advances in Neural Information Processing Systems. [S.l.]: Curran Associates, Inc., 2021, 34: 8780-8794.
- [102] SONG J, MENG C, ERMON S. Denoising diffusion implicit models[EB/OL]. (2020-10-06). <https://doi.org/10.48550/arXiv.2010.02502>.
- [103] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything[C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023: 3992-4003.
- [104] WU S, LI S, LI J, et al. LocateEdit-bench: A benchmark for instruction-based editing localization[EB/OL]. (2026-02-05). <https://doi.org/10.48550/arXiv.2602.05577>.
- [105] YU Z, NI J, LIN Y, et al. DiffForensics: Leveraging diffusion prior to image forgery detection and localization[C]//

- Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2024: 12765-12774.
- [106] NAM J H, MOON D H, LEE S C. M2sformer: Multi-spectral and multi-scale attention with edge-aware difficulty guidance for image forgery localization[C]//Proceedings of 2025 IEEE/CVF International Conference on Computer Vision (ICCV). Honolulu, HI, USA: IEEE, 2025: 15927-15938.
- [107] GUO X, LIU X, MASI I, et al. Language-guided hierarchical fine-grained image forgery detection and localization[J]. International Journal of Computer Vision, 2025, 133(5): 2670-2691.
- [108] SHA Z, LI Z, YU N, et al. DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models [C]//Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. [S.l.]: ACM, 2023: 3418-3432.
- [109] HE X, ZHOU Y, FAN B, et al. VLForgery face triad: Detection, localization and attribution via multimodal large language models[C]//Proceedings of Advances in Neural Information Processing Systems 38. San Diego, California, USA and Mexico City, Mexico: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2025: 180893-180927.
- [110] FERNANDEZ P, COUAIRON G, JÉGOU H, et al. The stable signature: Rooting watermarks in latent diffusion models [C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023: 22409-22420.
- [111] XU Q, MU W, LI J, et al. Advancements in AI-generated content forensics: A systematic literature review[J]. ACM Computing Surveys, 2026, 58(3): 1-36.
- [112] SUN Z, JIANG H, CHEN H, et al. ForgerySleuth: Empowering multimodal large language models for image manipulation detection[C]//Proceedings of Advances in Neural Information Processing Systems 38. San Diego, California, USA and Mexico City, Mexico: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2025: 169963-170002.
- [113] HUANG Z, HU J, LI X, et al. SIDA: Social media image deepfake detection, localization and explanation with large multimodal model[C]//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2025: 28831-28841.
- [114] HUANG Z, LI X, YANG X, et al. So-Fake: Benchmarking and explaining social media image forgery detection[EB/OL]. (2026-07-31). <https://doi.org/10.48550/arXiv.2505.18660>.
- [115] YU Y, ZHOU Y, LI B, et al. AgentFoX: LLM agent-guided fusion with explainability for AI-generated image detection[EB/OL]. (2026-07-03). <https://doi.org/10.48550/arXiv.2603.23115>.
- [116] REN S, YAO Y, ZEWEDE K, et al. Can multi-modal (reasoning) LLMs work as deepfake detectors?[EB/OL]. (2025-03-29). <https://doi.org/10.48550/arXiv.2503.20084>.
- [117] GUO X, SONG X, ZHANG Y, et al. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector[C]//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2025: 105-116.
- [118] GUO S, CUI J, HONG R. Rethinking VLMs for image forgery detection and localization[EB/OL]. (2026-03-13). <https://doi.org/10.48550/arXiv.2603.12930>.
- [119] UDAYANGA D, CHEN P Y, DAS P, et al. Can vision-language models reason about AI edits in images?[EB/OL]. (2026-07-30). <https://doi.org/10.48550/arXiv.2607.28464>.
- [120] LAI X, TIAN Z, CHEN Y, et al. LISA: Reasoning segmentation via large language model[C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2024: 9579-9589.
- [121] LAI Y, YU Z, WANG J, et al. ForensicZip: More tokens are better but not necessary in forensic vision-language models[EB/OL]. (2026-03-12). <https://doi.org/10.48550/arXiv.2603.12208>.
- [122] YAN Z, ZHANG Y, YUAN X, et al. DeepfakeBench: A comprehensive benchmark of deepfake detection[C]//Proceedings of Advances in Neural Information Processing Systems 36. New Orleans, Louisiana, USA: Neural Information Processing

Systems Foundation, Inc. (NeurIPS), 2023: 4534-4565.

- [123] LE T N, NGUYEN H H, YAMAGISHI J, et al. OpenForensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021: 10097-10107.
- [124] WU H, ZHOU J, TIAN J, et al. Robust image forgery detection over online social network shared images[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 13430-13439.
- [125] WU H, ZHOU J, TIAN J, et al. Robust image forgery detection against transmission over online social networks[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 443-456.
- [126] LI H, PENG R, LUO A, et al. Universal anti-forensics attack against image forgery detection via multi-modal guidance[EB/OL]. (2026-02-19). <https://doi.org/10.48550/arXiv.2602.06530>.

作者简介:



王伟(1990-),通信作者,男,教授,博士生导师,研究方向:计算机视觉和机器学习,E-mail: wei.wang@bjtu.edu.cn。



李昂(2001-),男,硕士研究生,研究方向:计算机视觉。



倪蓉蓉(1976-),女,教授,博士生导师,研究方向:数字图像取证、图像源及视频取证和数字水印技术及应用。



魏云超(1987-),男,教授,博士生导师,研究方向:计算机视觉和机器学习。



赵耀(1967-),男,教授,博士生导师,研究方向:数字媒体信息处理与智能分析、人工智能、计算机视觉、AIGC以及AI视频编码。

(编辑:张黄群)