文本图像篡改检测与定位综述

句福娇[1*],张 浩[1],齐光磊[2],王宏远[1]
(1. 北京工业大学计算机学院,北京 102101;2. 北京邮电大学世纪学院计算机科学与技术系,北京 102101)

摘 要: 随着生成式人工智能技术的快速发展,文本图像的篡改手段日趋智能和隐蔽,严重威胁学术诚信、信息安全与社会信任。文本图像篡改分析(检测与定位)旨在判别图像是否存在篡改,并进一步定位图像中被篡改的文本区域,以维护信息的真实性和图像的可信度。本文系统回顾了近年来该领域的研究进展,从单流视觉建模、多模态融合检测、文本语义与结构一致性分析三个视角梳理了现有的深度学习篡改分析方法,并分析各类方法的设计思路与适用场景。在此基础上,本文进一步从模型鲁棒性与工程部署两个横向维度,重点讨论了近年来出现的前沿技术,包括对抗样本训练策略、大型视觉语言预训练模型在文本一致性判定中的应用、跨语种与场景文本检测的挑战、面向嵌入式系统以实现高效部署的轻量化检测网络,以及融合语言模型生成解释以增强模型透明度和用户信任的可解释性方法。在评估基准方面,本文总结了现有公开数据集及其规模和特征,并对代表性方法的检测与定位性能和模型复杂度进行对比分析。最后,结合现有研究工作,本文提出了有待解决的难点与未来发展趋势,为文本图像篡改检测与定位领域提供了全面的技术视角和研究参考。

关键词: 文本图像篡改;篡改检测与定位;多模态融合;语义与结构一致性;对抗鲁棒性;视觉语言模型;可解释性取证

基金项目: 北京市自然科学基金(No.4242016,No.4244072)

中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2026)03-1364-27

电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250830

A Comprehensive Review of Text Image Tampering Detection and Localization

JU Fujiao[1*], ZHANG Hao[1], QI Guanglei[2], WANG Hongyuan[1]
(1. *College of Computer Science*, *Beijing University of Technology*, *Beijing* 102101, *China*;
2. *Department of Computer Science and Technology*, *Century College*, *Beijing University of Posts and Telecommunications*, *Beijing* 102101, *China*)

Abstract: With the rapid development of generative artificial intelligence, text images can be tampered with in increasingly subtle and realistic ways, posing severe threats to academic integrity, information security, and social trust. Text image tampering analysis, covering both tampering detection (image-level authenticity judgement) and tampering localization (pixel-level delineation of manipulated text regions), aims to verify image authenticity and provide fine-grained evidence for downstream forensics. This paper systematically reviews recent progress in this field and organizes deep learning-based methods from three perspectives: single-stream visual modeling for mining forensic traces, multimodal fusion for integrating complementary cues (e.g., spatial, frequency, and degradation artifacts), and semantic/structural consistency analysis for exploiting textual content and layout constraints. Beyond these methodological routes, we further highlight two cross-cutting dimensions that have gained momentum in recent years, namely robustness improvement under adversarial perturbations and real-world corruptions, and practical deployment including lightweight architectures and explainable outputs to enhance efficiency and user trust. We also discuss the emerging role of large pre-trained vision-language models (VLMs) in text consistency verification, as well as challenges in cross-language settings and in-the-wild scene text. For evaluation, we summarize publicly available datasets and commonly used metrics, and compare representative methods in terms of detection/localization performance and model complexity. Finally, we outline open problems and future research directions to facilitate further advances in text image tampering detection and localization.

Keywords: text image tampering; tampering detection and localization; multimodal fusion; semantic and structural consistency; adversarial robustness; vision-language models; explainable forensics

Foundation Item(s): Beijing Natural Science Foundation (No.4242016, No.4244072)

收稿日期:2025-09-23;录用日期:2026-02-28;责任编辑:覃怀银

*通信作者:句福娇

0 引言

随着数字化和人工智能技术的飞速发展,文本图像已经成为信息交流和文档认证中不可或缺的载体。例如,各类电子发票、合同扫描件、身份证图片以及带文字的屏幕截图等,都可以视为文本图像。文本图像篡改指的是对图片中承载文本语义的信息元素(包括文字字符、数字、符号以及印章等图文组合要素)进行恶意修改,以生成虚假内容或误导接收者的行为。与自然图像不同,即使文本图像中一个字符的轻微改变也可能导致信息含义的重大偏移,如在发票金额、合同条款或证书编号等关键信息上动手脚。这种精细篡改的隐蔽性极高,会给金融业务[1-2]、司法取证[3]、身份认证[4]等领域带来严重威胁。例如,近期上海一起伪造车证的案件中,犯罪分子利用AI工具在短时间内"生成"了20万张伪造行驶证,造成了巨大损失。为应对类似的新型欺诈,国内外已经开始着力开发能够检测并定位文本图像篡改的技术手段。

然而,文本图像篡改分析(检测与定位)面临独特挑战。与自然图像不同(图1左侧),文本图像往往具有均匀的背景和规则的排版结构,篡改区域可能仅为单个字符或文字框(图1右侧),与周围区域的差异极为细微,这使得传统依赖像素或纹理异变的篡改分析方法难以奏效[5]。其次,篡改后的图像通常会被进一步进行缩放、压缩、模糊等处理,以消除篡改痕迹,进一步增加了篡改分析难度。此外,文本图像篡改的语义影响远大于视觉变化(如发票金额"1"改为"7"),因此针对文本篡改的分析还需要结合语言一致性和内容逻辑进行判断。

面对以上挑战,近年来研究者逐渐从传统的手工特征分析转向基于深度学习的端到端方法。早期学者的主要工作是利用特殊的手工特征来检测篡改操作带来的扭曲[5]。一些文献[6-7]通过将外部打印件识别为篡改文本,将此任务转换为打印机源识别。另一些文献则通过研究字体[8]、文本行方向[9]、几何形状[10]、图像质量[11]、DCT系数[12]和局部纹理模式[13]的畸变来检测篡改文本。这些方法具有较高的可解释性,但在面对较为隐蔽的情况时,其泛化能力受到限制。最新的工作利用卷积神经网络(CNN)和Transformer架构,从纯视觉和多模态语义两大视角挖掘篡改线索[14-20]。典型工作包括采用双流网络融合图像的空间域与频域证据,以及结合图像特征与OCR提取的文本内容以发现细微的篡改痕迹。此外,引入注意力机制、多模态信息融合和对抗训练等技术进一步提升了模型对小范围篡改的定位准确率,使得这些方法在公开数据集上的性能相较早期手工特征方法取得了显著提升。

图1　自然图像与文本图像的差异性对比

Figure 1　Comparison between natural images and text images

为了明确本文的讨论范围,我们遵循图像取证领域的标准术语:"篡改检测(tampering detection)"指对图像是否被篡改进行二分类判断(图像级);"篡改定位(tampering localization)"指识别并分割出图像中被篡改的具体区域(像素级)。本文所综述的方法绝大多数聚焦于"篡改定位"任务,因为它提供了更丰富的证据信息,是更具挑战性的研究方向。在后续行文中,我们严格区分"检测"(图像级二分类判断)与"定位"(像素级篡改区域分割)两类任务:当不需区分二者时,统一使用"篡改分析(检测与定位)";当强调图像级判别时使用"篡改检测";当涉及像素级输出或篡改区域时统一使用"篡改定位"。

本文将对文本图像篡改分析(检测与定位)领域的研究现状进行全面回顾和总结。第1章介绍该领域常用的篡改数据集和评估指标,说明各类公开数据集的构建特点及主要性能指标的含义。第2章归纳了典型的文本篡改方式以及常见的后处理操作(如裁剪、模糊等)。第3章系统梳理了近年来基于深度学习的篡改分析方法:首先分别从单流视觉模型、多模态融合检测和文本语义与结构建模三种技术路线对

代表性方法进行分类综述，随后从对抗攻击与防御机制出发总结鲁棒训练策略，再从轻量化检测网络与模型可解释性实践两个工程维度讨论检测模型的高效部署与可信性增强，并在最后通过代表性方法在典型数据集上的性能与复杂度对比对上述方法进行综合分析。第4章深入讨论当前研究所面临的挑战和未来发展方向，包括多语言文本篡改检测、生成式AI带来的新威胁、模型实际部署问题以及法律伦理方面的考虑，并总结现有方法的局限。通过上述安排，本文旨在为文本图像篡改分析研究提供清晰的技术脉络和全面的参考指引。

1 数据集与评价指标

在数字化转型浪潮下，智能文档处理系统已广泛应用于电子合同签署、自动化票据核验等关键场景。这类系统依赖OCR识别、文档结构解析等技术处理文本图像，来实现业务流程智能化，却也面临文本图像篡改带来的身份冒用、财务欺诈等新型风险[21-24]。

自2020年以来，随着文本生成模型（如基于Transformer的布局生成[25]、Diffusion驱动的笔迹合成[26]）的突破性进展，篡改手段已从传统的局部篡改（如金额涂改）扩展到全图像生成（如虚拟发票创建）、语义级篡改（如合同条款替换）等复杂形态。为应对这一威胁，构建专业化的文本图像篡改数据集成为研究重点。

如图2所示，该图展示了若干典型的文本图像，包括证书类文档以及结构化表单。这些样本既包含扫描获得的图像，也包含拍摄/截图得到的图像，从而体现了文本图像在采集链路、版式结构、文本密度与图文混排方式上的多样性，以及潜在的篡改风险点（如数字篡改、印章篡改等）。在实际应用中，同一文档往往同时具备表单、表格与图形等结构特征，且其采集方式（扫描/拍摄/截图）也可与文档类型交叉组合。定期更新包含多样化篡改内容的基准数据集，并确保检测模型针对代表性的篡改技术进行测试，是提升算法鲁棒性的关键。当前，每个基准数据集或相关研究均通过标准化评估指标验证其可靠性，从而为篡改检测与定位技术的迭代优化提供可比依据。

1.1 基准数据集

图像操纵检测（Image Manipulation Detection，IMD）作为图像取证的热点话题，随着数据集的不断增加，研究也越来越深入。篡改的目标通常是一个明显的语义对象，如图1左侧图像中拼接的两个人。相比之下，文本篡改旨在扭曲所携带的上下文信息。被篡改的对象，即文本，通常与文档的其余部分保持相似的视觉语义和外观[5]。

然而，构建一个文本篡改数据集并非易事。为确保语义与外观的一致性，往往需要专业人员进行高质量的手动文本篡改操作，并且文本篡改和细粒度注释的成本昂贵[5]。因此，文本篡改数据集数量稀少。为便于后续方法在相同条件下进行定量比较，本文选取近年来被广泛采用的若干代表性文本图像篡改数据集，从规模、篡改方式、后处理操作以及语种、场景和标注粒度等多个维度进行归纳和对比，如表1和表2所示。其中，表1~表2所列“公开数据集”的获取链接已在尾注中列出，便于读者下载与复现。

表1统计了各数据集的样本规模（包括图像数量以及篡改/未篡改样本的构成）、构建年份、支持的篡改方式以及是否包含多种后处理操作。从中可以看出，早期数据集（如Find it!、SACP等）规模相对较小，主要采用少量基本的篡改操作，后处理手段也较为有限；随着竞赛型和大规模文档数据集的出现，后续数据集在样本数量、篡改方式多样性以及恶意后处理（如压缩、模糊、截图捕获和图像混合等）方面均有显著提升，更贴近真实攻击场景的复杂性。

仅从样本数量和篡改方式的角度尚不足以全面刻画文本图像篡改数据集的差异。在实际应用中，语种、场景/文档类型以及标注粒度等因素同样会显著影响模型的设计与评估结果。因此，表2进一步从这些维度对上述数据集进行了归纳比较。

扫描图表 收据 证书 包装说明 申请表 图表

图2 文本图像的数据种类

Figure 2 Data categories of text images

表1　文本图像篡改数据集统计信息(规模/篡改方式/后处理)

Table 1　Statistics of text image tampering datasets (scale/tampering types/post-processing)

数据集	统计数字			年份	篡改方式						有无后处理	
	真实图片	篡改图片	总图片		复制移动	拼接	生成	覆盖	修复	删除		
Find it![27]	940	240	1 180	2018	√	√	√	√	√	√	×	—
SACP[28]	—	—	2 005	2020	—	—	—	—	—	—	×	—
DID*[29]	—	—	—	2021	√	√				√	√	缩放、对比度/亮度调节、边界模糊
T-SROIE[3]	0	986	986	2022			√				×	—
T-IC13[30]	84	378	462	2022			√				√	photoshop
RIFLC[31]	0	8 000	8 000	2022	—	—	—	—	—	—	√	缩放、压缩、裁剪、截图捕获与社交传输
TTI[32]	3 006	15 994	19 000	2023	—	—	—	—	—	—	×	—
DocTamper[33]	0	170 000	170 000	2023	√	√	√				√	—
Find it again[34]	825	163	988	2023	√	√	√	√	√	√	√	人工校正标注
TextTamper[35]	—	—	49 500	2024		√					√	压缩、截图捕获、边界模糊、噪声和图像混合
RTM[5]	3 000	6 000	9 000	2025	√	√	√	√	√		√	缩放、压缩

注:*表示该数据集未开放获取;—表示文献中未明确说明。

表2　文本图像篡改数据集属性对比(语种/场景/标注粒度)

Table 2　Comparison of text image tampering dataset attributes (language/scenario/annotation granularity)

数据集	语种	场景/文档类型	标注粒度	关键特征/演进
Find it![27]	法语	收据、票据	像素级+文本级(篡改文本标注)	早期法语票据场景数据集,采用人工手工篡改,规模较小但提供精细像素级与文本级标注
SACP[28]	中英双语	证书图像	像素级	安全AI挑战者计划竞赛数据集,聚焦证书篡改检测,包含多种对抗性攻击与后处理方式
DID[29]	未说明	通用文本图像(手机拍摄)	像素级	由多款手机拍摄的高分辨率文本图像构成,考虑缩放、对比度/亮度调整、边界模糊等多种后处理,但数据集本身未开放获取
T-SROIE[3]	英语	收据、票据(基于SROIE)	像素级	在SROIE票据数据集上合成生成式文本篡改,面向票据场景,提供训练/验证/测试划分,文本以数字和英文字符为主
T-IC13[30]	英语	场景文本图像(基于IC13)	像素级	将生成式文本篡改扩展到自然场景文本图像,背景纹理复杂,包含多种字体和场景,对模型判别能力要求更高
RIFLC[31]	中英双语	证书、官方文书、截图、门脸图等多类型文档	像素级	世界图像伪造定位挑战赛数据集,涵盖多种文档类型和复杂后处理,模拟真实世界多源伪造场景
TTI[32]	中英双语	电商场景文本图像	像素级	大规模电商场景文本图像数据集,包含电商平台常见的多种图像编辑与压缩操作,贴近实际业务场景
DocTamper[33]	中英双语	合同、发票、收据等现实文档	像素级	17万张中英双语合成篡改样本,涵盖复制-移动、拼接、生成式等多种篡改方式,并设置两个跨域测试子集以评估泛化能力
Find it again[34]	英语	收据图像(基于SROIE)	篡改区域级+语义实体级多层级标注	基于公开SROIE收据图像构建,采用人工篡改生成样本,同时提供图像级、篡改区域级以及语义实体级标注,支持多种篡改类型与实体类别评测
TextTamper[35]	中英双语	通用文本图像	像素级	通过拼接与生成式方法大规模合成篡改文本图像,引入图像混合等恶意后处理操作,强调对复杂后处理和语义一致性的鲁棒性
RTM[5]	中英双语	真实采集文本图像(电商、志愿者拍摄等)	像素级	面向现实中文场景的手工篡改数据集,图像来源多样(电商脱敏数据、志愿者拍摄、开源数据集等),采用精细像素级掩码标注,突出真实世界难度

表2从语种、场景/文档类型、标注粒度和关键特征/演进等方面对典型数据集进行了对比。可以看到,Find it!、T-SROIE等数据集主要面向票据类文档场景,语种以法语或英语为主,并提供像素级或区域级的篡改位置标注;RIFLC和TTI等竞赛数据集则覆盖了证书、合同等更多样化的文本图像,在后处理和

跨设备采集方面具有更高的多样性；DocTamper通过合成方式构建了大规模中英双语文档篡改数据集，并设置跨域测试子集，强调对模型泛化能力的评估；TextTamper和RTM等最新数据集则进一步聚焦真实业务场景，前者首次系统引入图像混合作为恶意后处理手段，后者则针对现实中文档采集，由专业人员完成多种篡改操作并提供精细的像素级掩码标注，显著提升了真实场景下的评估难度。

综合表1和表2可以看出，文本图像篡改数据集在近几年呈现出由“小规模、单一场景、简单篡改”向“大规模、多场景、多语种、复杂篡改与多样后处理”演进的趋势。早期数据集主要用于验证文本图像篡改检测与定位方法的基本可行性，在语种、场景和标注粒度上相对受限；随着多模态取证需求的提升，新近数据集在跨语种、真实文档采集、复杂后处理和跨域测试等方面不断扩展，为后续方法在不同应用场景下的鲁棒性和泛化能力研究提供了更具挑战性的基准环境。

1.2 评价指标

文本图像篡改定位任务本质上可以建模为像素级二分类问题，即将每个像素划分为“篡改”与“未篡改”两类。因此，可以通过混淆矩阵中的真阳性（True Positive，TP）、真阴性（True Negative，TN）、假阳性（False Positive，FP）和假阴性（False Negative，FN）四种情况来刻画模型的预测结果，并在此基础上构造相应的评价指标。其中，TP表示被正确预测为篡改像素的篡改像素，TN表示被正确预测为未篡改像素的未篡改像素，FP表示将未篡改像素错误预测为篡改像素，FN表示将篡改像素错误预测为未篡改像素。

在文本图像篡改定位研究中，常用的像素级指标包括精确率（precision）、召回率（recall）、F1分数（F1-score）、交并比（Intersection over Union，IoU）、ROC曲线下面积（Area Under the Curve，AUC）、假阴性率（False Negative Rate，FNR）以及Matthews相关系数（MCC）[36]。其中，精确率和召回率可以由混淆矩阵直接计算得到，进一步可获得F1-score等综合指标；IoU主要用于衡量预测掩码与真实掩码之间的重叠程度；AUC、FNR和MCC则从阈值无关性能、漏检风险以及全局相关性等不同角度对模型表现进行补充刻画。上述指标的数学定义分别如式（1）~（6）所示。

$$P_{\text{recision}} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{1}$$

$$R_{\text{ecall}} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{2}$$

$$\text{F1-score} = 2 \times \frac{P_{\text{recision}} \times R_{\text{ecall}}}{P_{\text{recision}} + R_{\text{ecall}}} \tag{3}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}} \tag{4}$$

$$\text{FNR} = \frac{\text{FN}}{\text{FN}+\text{TP}} \tag{5}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{6}$$

虽然这些指标在形式上都源于像素级混淆矩阵，但它们关注的侧重点和适用场景并不相同。为突出各指标在文本图像篡改定位任务中的作用与差异，下面对其主要功能和典型使用场景进行简要概括。

综合上述指标可以看出，文本图像篡改定位任务中常用的评价指标大致可以分为三类：其一是以精确率、召回率和F1-score为代表的像素级分类性能指标，侧重衡量模型在“篡改/未篡改”像素区分上的准确性与完整性；其二是以IoU为代表的空间重叠指标，用于评估预测篡改掩码与真实标注之间的重合程度，更关注篡改区域的定位质量；其三是AUC、FNR和MCC等风险与全局相关性指标，前者从阈值无关的角度刻画整体区分能力，后两者分别强调像素级漏检风险和在类别不平衡条件下的综合表现。在实际研究中，通常会综合报告F1-score和IoU以全面反映像素级定位性能，并在高风险应用场景下引入FNR或MCC等指标，以更好地刻画漏检代价与整体可靠性。

2 篡改及后处理方式

在文本图像篡改定位任务中，篡改操作的多样性与后处理技术的复杂性共同构成了定位算法面临的主要挑战。为了生成具有高度视觉一致性的篡改图像，攻击者通常采用多种篡改方式对图像中的文本内容进行修改，并辅以缩放、压缩、模糊等图像处理操作以掩盖篡改痕迹，削弱定位模型对异常区域的响应能力[35]。因此，系统梳理文本图像中的典型篡改类型及其后处理手段，对于构建具备广泛适应性和鲁棒

①表1~表2中公开数据集的获取链接如下：

1.Find it!：http://findit.univ-lr.fr/download-the-dataset/

2. SACP：https://tianchi. aliyun. com/competition/entrance/531812/introduction

3.T-SROIE：https://github.com/wangyuxin87/Tampered_sroie

4.T-IC13：https://github.com/wangyuxin87/Tampered-IC13

5. RIFLC：https://tianchi. aliyun. com/competition/entrance/531945/introduction

6.TTI：https://tianchi.aliyun.com/competition/entrance/532052/introduction

7.DocTamper：https://github.com/qcf-568/DocTamper?tab=readme-ov-file.

8.Find it again：http://l3i-share.univ-lr.fr/2023Finditagain

9.TextTamper：https://github.com/nbudongli/text-image-forgery-detection

10.RTM：https://github.com/DrLuo/RTM

性的篡改分析方法(检测与定位)具有重要意义。

2.1 文本图像篡改方式

数据集通常是为了模拟真实世界的文本篡改,通过观察真实应用场景可发现,攻击者通常对收集的文本图像进行精细的手动篡改。针对此类篡改行为,本文归纳了几种典型的实现技术手段[5]。

(1)复制后移动(copy-move):操作者复制文本区域(随机选择的字符、单词或句子),粘贴到同一图像上,而不破坏布局。如图3(a)所示,将黄线表示的文本区域复制粘贴到精心选择的位置。由于复制的文本区域的方向、字体、大小和颜色与周围的文本一致,因此图像的整体结构保持得很好,增加了文本操作检测的难度。

(2)拼接(splicing):将源图像中的文本区域复制粘贴到目标图像中。这是现实世界中常见的文本操作。如图3(b)所示,从源文件中复制手写签名粘贴到目标文件中伪造认证。

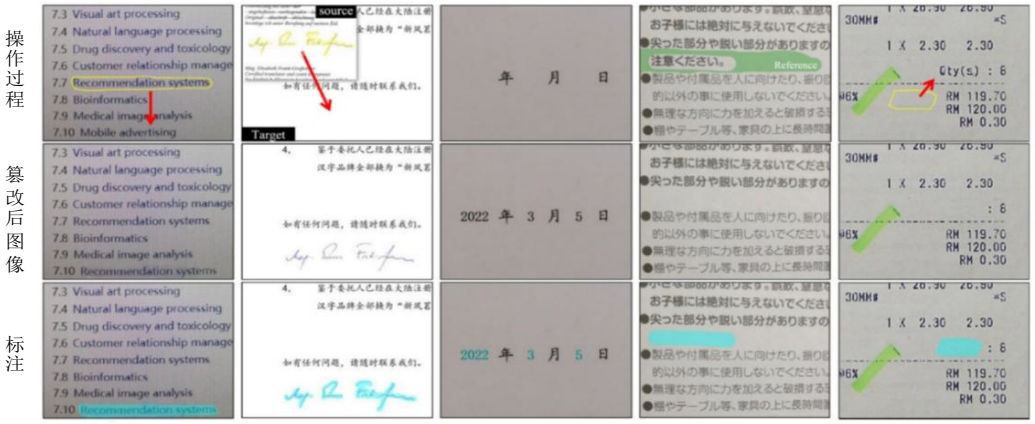
(3)插入(insertion):操作员在图像上插入新的文本,例如一个单词或句子。如图3(c)所示,添加的文本与周围的文本颜色和字体相似。

(4)修复(inpainting):通过工具PS或算法智能填补文本区域,一般参考周围像素内容。如图3(d)所示,根据参考背景的图案,即绿色区域,对这些像素进行填充,去除文本区域。为了使图像自然无破绽,操作人员被要求仔细选择排除其他文本的参考区域。

(5)覆盖(coverage):操作者通过用图像中相似区域手动覆盖目标内容,如图3(e)所示。通常人为选择相似区域(如背景)进行复制粘贴遮盖,快速遮盖文字,强调"遮住"而非"填补"。

(6)生成式篡改(generative manipulation):操作者训练生成模型(如SRNet[23])合成新的文本字符,以替代原始内容或插入篡改信息。原始图像中待篡改区域的字符,被生成模型所生成的字符替换。该字符的字体、颜色、排版与周围文本高度匹配,难以通过传统的边缘分析或纹理不一致性检测手段识别。

(7)删除(deletion):操作者一般通过图像编辑工具将图像中的特定文本区域移除,达到抹除信息的目的。通常指将文字区域"清除"掉,结果可以为空白、模糊或被填补。



(a) 复制后移动　(b) 拼接　(c) 插入　(d) 修复　(e) 覆盖

(a) Copy-move　(b) Splicing　(c) Insertion　(d) Inpainting　(e) Coverage

图3　几种典型的文本图像篡改方式

Figure 3　Typical types of text image tampering

除上述这种常见的篡改方式之外,结合这些基本操作可以创建更为复杂的篡改方式。例如:先拼接异源文本区域,再对边缘区域进行修复或模糊处理可增强篡改的视觉一致性;从图像中删除原始文本,然后插入新内容可以有效地实现文本替换等等。在实际应用中,通常表现为:插入签名、篡改文本、修改价格等等。

2.2 图像后处理方式

在现实场景中,文本图像经社交网络传输后通常会经历仿射变换、图像重捕获和有损压缩等处理,导致潜在的篡改痕迹失真。除此之外,攻击者还可能采用复杂的图像混合手段进一步模糊篡改区域,以增强篡改图像的整体协调性和可信度,从而增加检测难度。本节将介绍几种经典的后处理操作方式供研究者参考。

(1)缩放(scaling):操作者在完成图像篡改后,对整幅图像或其中某一部分进行放大或缩小。图像经过缩放处理后,篡改区域的像素结构被重排,使篡改痕迹不易通过图像层级的特征被检测。由于缩放操作保留了原有的布局比例,因此图像整体结构未被

破坏。

(2)亮度/对比度调节(brightness/contrast adjustment):通过调整图像的亮度或对比度,操作者增强篡改区域与背景的一致性。篡改区域在处理前颜色偏暗,与周围区域存在差异,通过对整图亮度提升,使该差异被有效削弱,增强图像的视觉自然性。

(3)边界模糊(edge blurring):为了消除剪贴操作产生的生硬边缘,操作者在篡改区域边界处施加模糊处理。原始粘贴区域边缘清晰,而应用模糊后边界过渡更为自然,降低了肉眼和检测算法识别的可能性。

(4)压缩(compression):在篡改完成后,图像被保存为JPEG等有损压缩格式。压缩过程中引入的伪影和噪声掩盖了篡改痕迹,使得边缘不连续、色彩不一致等异常特征被压缩噪声干扰,从而影响算法检测。

(5)裁剪(cropping):通过对图像的局部裁剪,操作者可以有意删除带有篡改区域的部分。图像底部原含有篡改内容,经裁剪后该区域被完全移除,仅保留可信区域,意图规避检测与定位分析。

(6)截图捕获与社交传输(screenshot and social transmission):将篡改后的图像通过截图工具截取并经由社交平台进行转发,该过程会导致图像被重新编码、尺寸变化或压缩,增加算法检测复杂度,同时社交平台的自动处理流程可能破坏原始篡改痕迹。

(7)噪声添加(noise injection):在图像中叠加噪声以增加内容一致性。原始图像中篡改区域边界过于光滑,通过添加高斯噪声,图像整体显得更加"粗糙"自然,影响篡改检测算法对边缘异常的识别。

(8)图像混合(image blending):操作者利用图像混合技术(如泊松融合[22]或深度图像混合[37])使篡改区域适应目标图像,以协调篡改图像,使篡改痕迹难以察觉。篡改后区域与背景之间存在差异,通过融合后,色彩、纹理过渡自然,极大增强了图像的真实感和一致性。

3 文本图像篡改检测与定位方法

随着文本图像篡改技术的不断发展,篡改行为呈现出更高的隐蔽性与多样性,给篡改定位任务带来了巨大挑战。近年来,研究者提出了多种基于深度学习的篡改定位方法,试图从不同视角对文本图像中的篡改区域进行精准识别与定位。相比传统依赖手工特征的技术路径,深度模型能够从大规模数据中自动学习文本篡改的潜在特征,展现出更强的判别能力与泛化性能。为应对文本图像在结构复杂性、篡改类型多样性以及后处理策略多变性等方面带来的挑战。

近年来研究者提出了多种篡改定位框架(如图4所示),涵盖从单流网络到双流、多模态架构,从低层视觉线索到高层语义建模的多种技术路线。本章将对当前具有代表性的文本图像篡改定位分析方法(检测与定位)进行系统归类与梳理。总体来看,这些方法可根据输入特征的类型与网络结构的设计进行归类,主要包括:基于单通路视觉建模的方法、基于多模态融合的方法,以及基于文本语义与结构建模的方法。

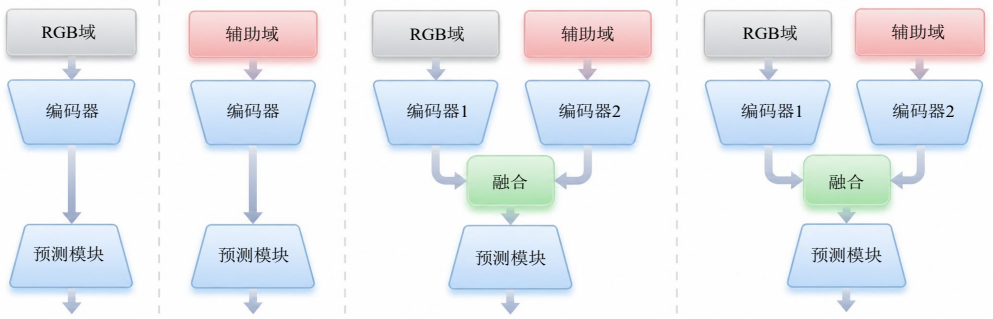


图4 主流的文本图像篡改定位框架

Figure 4 Mainstream frameworks for text image tampering localization

3.1 基于视觉特征的单流架构方法

在文本图像篡改定位的早期研究中,大多数方法依赖于单一信息源进行篡改区域建模,主要从图像的像素层级出发,利用视觉特征(如边缘纹理、颜色异常、压缩伪影等)进行篡改区域的检测与定位。这类方法通常采用编码-解码结构,通过卷积神经网络提取多层次特征,并引入注意力机制、上下文聚合模块等增强模型对细粒度篡改线索的感知能力。尽管结构较为简洁,但在多个公开数据集的受控条件下仍取得了较为稳定的定位性能,然而由于主要依赖单一视觉域线索,其在复杂后处理以及真实场景中"弱痕迹、小区域"篡改的条件下的鲁棒性仍相对有限,因此更适用于图像质量较好且后处理强度有限的常规场景。下面将介绍几种具有代表性的单流架构方法,以揭示其设计思路及性能特点。

Liang等人[14]较早关注到图像混合(image blend-

ing)等恶意后处理会显著削弱文本篡改边界痕迹,提出了一种抗混合干扰的单流网络。该方法基于编码-解码框架,在特征提取过程中通过全局与局部特征增强模块突出潜在篡改区域并强化细粒度边缘纹理,从而在作者构建的混合文本图像数据集及SACP数据集[28]上取得了较好的F1-score和IoU,验证了单流视觉网络在复杂后处理场景下的可行性。

在此基础上,同一研究团队进一步提出了TIFDM网络[35],其整体结构如图5所示。TIFDM仍采用单流编码-解码框架,但在特征建模上进行了更有针对性的设计:篡改痕迹融合模块FTFN从空间域特征与误差域特征中提取互补线索并进行融合,以恢复在JPEG压缩、重采样等操作下被弱化的篡改痕迹;低高层特征融合模块LHSE通过整合浅层细节与高层语义信息,兼顾小目标边缘与全局上下文;文本伪造注意力模块TFAM则在通道维度上显式区分文本区域与背景区域,抑制与篡改无关的冗余响应。实验结果表明,TIFDM在TextTamper[35]、SACP[28]、TTI[32]和RIFLC[31]等多个数据集上均取得了领先的F1-score和IoU,在模糊、亮度变化、压缩等多种后处理条件下性能下降较小,体现了面向弱痕迹场景的单流视觉模型的代表性进展。

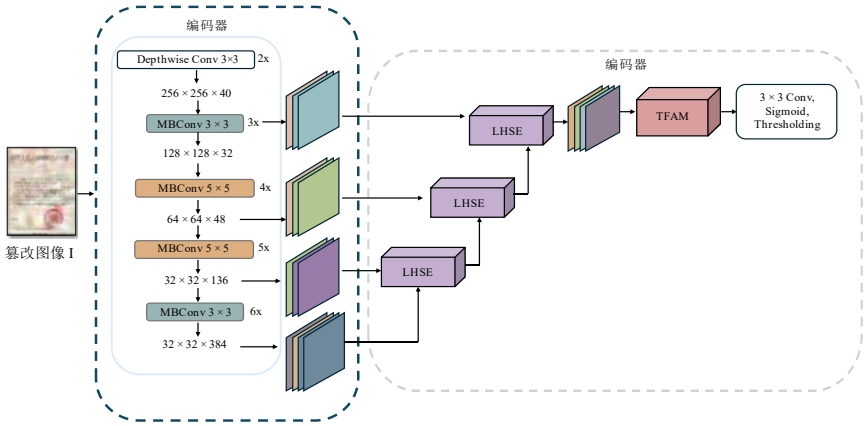


图5　TIFDM网络架构

Figure 5　Architecture of the TIFDM

针对证书等文档场景中篡改区域尺度多样、版式复杂的问题,Sun等人[38]提出了多级特征注意力网络MFAN,其框架如图6所示。MFAN以ResNet为主干网络,在编码阶段提取多尺度语义特征,并设计全局通道注意力与局部空间注意力相结合的特征重标定模块,一方面突出与文字、印章等关键区域相关的通道响应,另一方面抑制背景纹理和复杂版式带来的干扰。解码阶段通过级联的卷积与上采样模块逐步恢复空间分辨率,并利用跳跃连接融合浅层细节信息,最终输出像素级篡改掩码。在SACP与RIFLC等证书数据集上的实验表明,MFAN在IoU和F1-score等指标上明显优于CFA[39]、Mantra-Net[40]等通用篡改定位方法;在多种自然图像篡改数据集上的结果也显示,该模型在跨场景迁移时仍具有较好的泛化能力,说明在单流框架下通过多级注意力对文档场景进行定制化建模是行之有效的。

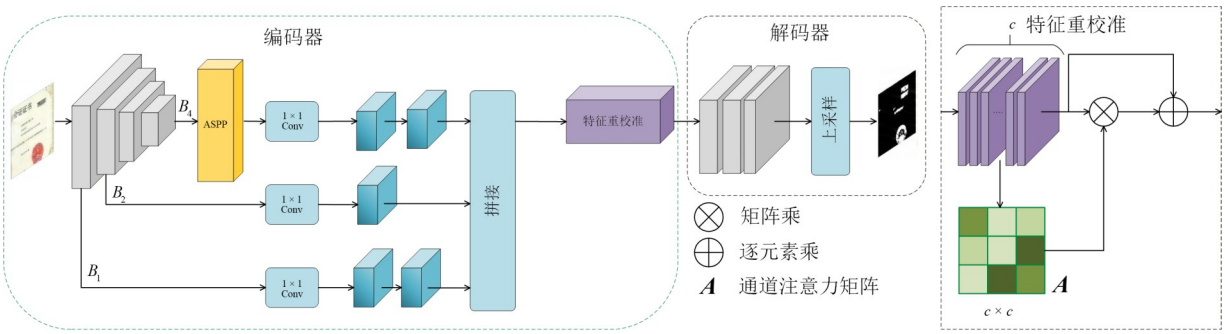


图6　多级特征注意力网络架构

Figure 6　Architecture of the multi-level feature attention network

此外，通用图像篡改定位领域的一些单流网络设计也可为文本图像篡改定位提供可迁移的结构组件。例如，Zhuang 等人[19]提出的 Dense-FCN 以密集连接增强特征复用，并结合空洞卷积扩大有效感受野，从而在单流 FCN 框架下兼顾细粒度边界刻画与上下文聚合能力。对文本图像而言，这类设计的借鉴价值主要体现在两点：其一，密集连接更有利于保留字符笔画、局部纹理与细窄边界等浅层信息，降低小范围篡改在逐层下采样过程中被“抹平”的风险；其二，空洞卷积提供了更大的上下文视野，有助于建模跨字符/跨字段的局部一致性，从而辅助区分“正常版面规律变化”与“篡改导致的不一致”。

总体来看，以 Liang 等人及其后续 TIFDM 工作为代表的单流视觉方法，通过引入误差域特征融合并针对图像混合、压缩等操作进行专门设计，在一定程度上缓解了后处理带来的弱痕迹问题；MFAN 则结合多级注意力机制，提升了在证书等复杂文档场景下对篡改区域的建模能力，说明在充分挖掘视觉线索时，单流架构仍具有较强的表现力。从机理上看，这类方法完全工作在视觉域，依托卷积网络和注意力模块在多尺度感受野上自动学习边缘、纹理与成像噪声统计的局部不一致，能够较好刻画复制移动、拼接和图像混合等操作打破文档背景一致性的效应，因此在图像质量较好、压缩和模糊程度有限的标准票据和证书场景中往往可以取得稳定表现。但也正因为所有信息都沿着单一视觉通路传播，对 RGB、误差域、频域等不同视觉视角之间的互补线索利用有限，当篡改经过强压缩、重采样等复杂后处理，或主要体现在金额数字等语义层面时，单流模型的判别边界容易变得脆弱，漏检风险明显增大。相比之下，后续将要介绍的多通路融合方法可以通过频域和残差域补偿弱纹理线索，语义与结构建模方法则利用文本内容和版式先验发现逻辑矛盾，从多层次弥补单流视觉路线的局限。为进一步缓解上述问题，后续工作开始在视觉模态内部引入多分支特征提取与多通路融合，通过并行建模不同视觉域来增强篡改痕迹的感知能力，这也为下一节的多流架构方法奠定了基础。

3.2 基于多通路融合的多流架构方法

多通路融合方法通过构建双流或多流网络架构，从图像的不同表现域（如 RGB 域、频率域、残差域）或不同模态（如视觉与 OCR、DCT、ELA 等）提取篡改线索，并在特征层或决策层进行融合，增强对篡改痕迹的感知能力，以实现篡改区域的高鲁棒、高精度识别。这类方法充分考虑了不同模态线索之间的互补性，通过互补增强篡改线索，有效弥补了单一模态信息在表征能力上的局限，在复杂干扰场景下展现出更强的泛化性能与定位稳定性。本节将系统介绍当前主流的多模态篡改定位方法，分析其特征融合策略及在实际数据集上的表现。

早期的多流方法通常采用两条并行分支以提取不同域的特征。图 7 展示了 Xu 等人[29]提出的双流网络针对文本图像篡改，通过 Inception 模块构建的空间信息提取分支（SIEN）捕捉不自然边界与对比度不一致等视觉痕迹，同时通过六个残差滤波器构建的像素相关性网络（CFEN）提取缩放、复制粘贴引起的异常像素相关性。这两路特征在判别网络中进行融合后对每个图像补丁是否篡改进行分类，实现精细的区域级定位。实验证明，该方法在自建 DID 数据集上取得 Recall 为 0.99、IoU 为 0.85 等优异性能；且在任务更复杂的 SACP 数据集上仍优于其他基线方法（Dense-FCN[19]、ManTra-Net[40]、HLED[41]）。特别是在 JPEG 压缩与高斯噪声扰动下，该模型的 F1-score 保持在 0.88 以上，展现出出色的鲁棒性与跨场景泛化能力。

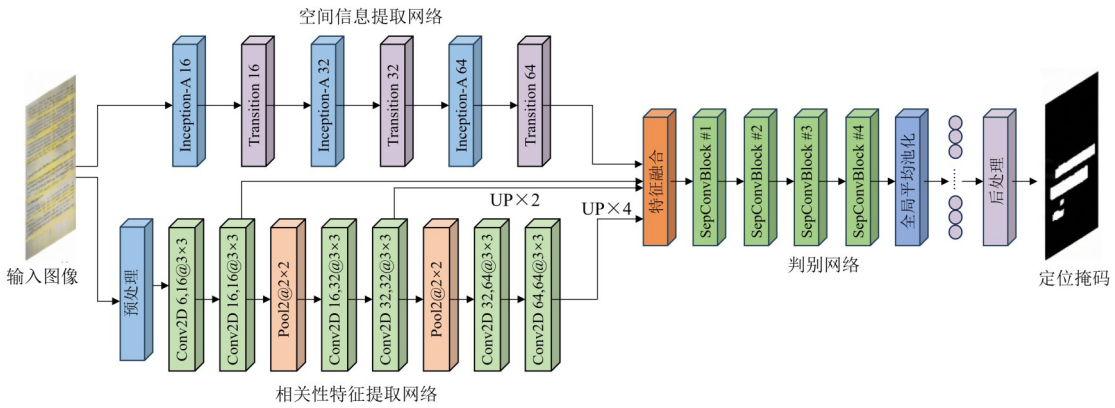


图 7 Xu 等人提出的区域级文本篡改定位模型架构

Figure 7 Architecture of the region-level text tampering localization model proposed by Xu et al

在文档图像场景中,也有工作尝试从更细粒度的字符纹理层面构建多流定位框架。Liao 等人[42]提出的 CTP-Net(Character Texture Perception Network)就是代表之一。该方法同样以单张文档图像为输入,但通过光学字符识别(OCR)结果将图像拆分为两个互补视图:一条字符纹理通路(Character Texture Stream,CTS)专注于文字区域,提取字符笔画、轮廓锐度等细粒度纹理特征;另一条整图纹理通路(Image Texture Stream,ITS)则在整幅图像上建模版式结构和背景纹理分布。两条通路在后续篡改定位子网络中逐级融合,既能利用文档整体排版与纸张纹理作为先验,又能放大被篡改字符在局部纹理上的不一致性。为缓解真实篡改样本匮乏的问题,作者还设计了合成伪造样本的 FCTM 数据集,实验表明 CTP-Net 在该数据集以及 SACP、DocTamper 等基准上均取得了优于多种单流方法的定位精度,尤其在字符级小尺度篡改区域上具有更好的定位能力。而在通用细粒度识别研究中也有工作通过融合全局与局部视角,并借助注意力机制过滤冗余、聚焦关键区域以提升细节辨识能力[43];这一"全局结构—局部细节"协同建模思路可为文本图像篡改检测与定位中的多视角线索融合提供借鉴。

随后,研究者将注意力引向不同模态的融合。Qu 等人[33]提出了一种多模态 Transformer 框架 DTD(Document Tampering Detector),如图 8 所示。该方法使用了视觉感知与频域感知双分支,视觉感知模块采用多个卷积块处理原始图像以获取边缘与结构信息,而频域感知模块则基于 DCT 系数与量化表建模,补偿视觉域在压缩或模糊条件下的表达缺陷。随后通过通道注意力将两路特征融合后送入 Swin Transformer 编码器[44]。该模型还采用类似多视角迭代解码器的策略分步识别篡改区域,并引入了"从易到难"的学习策略 CLTD(Curriculum Learning for Tampering Detection)以增强对压缩失真的鲁棒性。DTD 在自建的 DocTamper 数据集(含 17 万张中英文合成图像)及 T-SROIE 票据[3]篡改数据集上性能优异:在 DocTamper 上 F1-score 达 0.792,较次优模型(CAT-Net[45])提升 9.2%;在两个跨域测试子集(DocTamper-FCD 与 DocTamper-SCD)上,F1-score 分别达到 0.816 与 0.754,相较次优模型提升 26.3% 与 12.3%。此外,在 T-SROIE 数据集上达到 0.992 7 的 F1-score,刷新了该领域现有最高水平。这些结果表明,频域模态的引入有效补偿了视觉模态在压缩模糊条件下的信息缺失,增强了定位准确性。

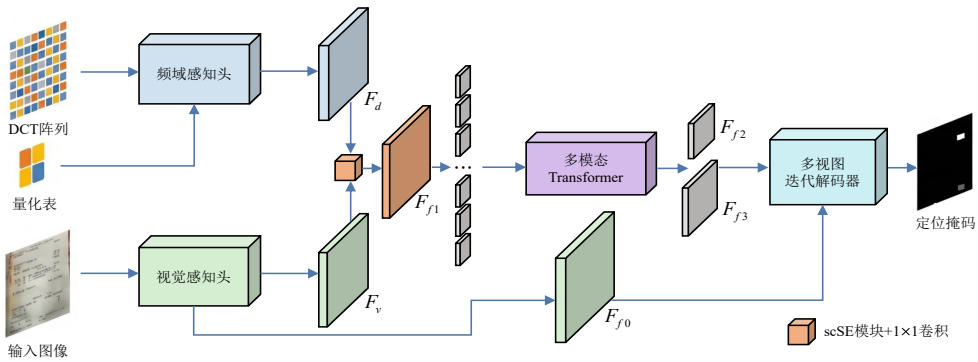


图 8　DTD 模型架构

Figure 8　Architecture of the DTD

在引入 DCT 频域分支并取得显著性能提升之后,Chen 等人[46]进一步提出了 FFDN(Frequency Feature fusion and Decomposition Network),从特征融合和频率分解两个层面细化对频域信息的利用。FFDN 延续了"视觉分支+频域分支"的双流结构,同时设计了视觉增强模块 VEM(Visual Enhancement Module)和类小波频率增强模块,WFE(Wavelet-like Frequency Enhancement)。其中,VEM 通过零初始化卷积等设计,在保持原始 RGB 语义特征完整性的前提下,将频域线索注入视觉分支,显式放大篡改区域的细微纹理差异;WFE 则对特征进行多尺度频率分解,将高低频子带分别建模并进行聚合,以保留在下采样过程中容易丢失的高频细节,使模型在处理小尺寸篡改区域时更加敏感。实验结果显示,在 DocTamper 及其高压缩、跨域划分上,FFDN 相比 DTD 在 F1-score 和 IoU 等指标上均有明显提升,尤其在强 JPEG 压缩和多次社交平台转发等复杂场景中仍能稳定发现篡改区域,进一步验证了频域增强策略在高压缩场景下的有效性。

与此同时,一些工作开始利用 OCR 文本信息融合视觉特征。例如,Yu 等人[47]针对手机截图中文本字符边界模糊的问题,提出了 STFL-Net 架构。该网络采用 RGB 分支提取整体视觉信息,并通过 PP-OCR 工具[48]识别的文本区域引导另一条 OCR 分支,使网络更聚焦于潜在的文本篡改区域。两条流在编码阶段通过双向跨模态注意力 DCMA(Dual Cross-Modal Attention)融合特征,在解码阶段通过空间通道注意

力 SCSE(Spatial Channel Squeeze-and-Excitation)强化重要特征响应。此外,作者引入多教师知识蒸馏策略,在预训练阶段加入多样本化篡改样本,加固了模型面对不同篡改方式时的泛化能力。在作者自建的 STFD 截图篡改数据集上,STFL-Net 在多种后处理(压缩、模糊、缩放)下均取得了明显优于主流方法(MVSS-Net[16]、DFCN[19]、Mantra-Net[40]等),尤其针对真实应用场景的定位任务表现更加稳定,证明了 OCR 导向多流融合对边界模糊场景的有效性。

更进一步,Luo 等人[5]面向实际应用场景提出了非对称双流架构的 ASC-Former,如图 9 所示,该方法由 RGB 主干流和多模态辅助流组成:辅助流同时提取频域、SRM 残差域、ELA 误差域等多种篡改线索,并通过一致性聚合中心 CA Hub(Consistency-aware Aggregation Hub)模块动态选取最具判别力的信息,再通过门控领域注意力融合 GCNF(Gated Cross Neighborhood-attention Fusion)模块实现多尺度跨模态特征交叉融合。训练阶段加入了篡改与真实对比学习 TAC(Tampered-Authentic Contrastive learning)模块,以强化模型区分真伪区域的能力。在自建的 RTM 真实篡改数据集、T-SROIE 票据[3]和 T-IC13[30]场景文本数据集上,ASC-Former 在像素级 F1-score 和 IoU 均显著高于先进方法,在小区域和微弱痕迹的篡改场景中表现尤为出色。此外,该框架具备良好的可扩展性,可灵活整合更多模态信息。

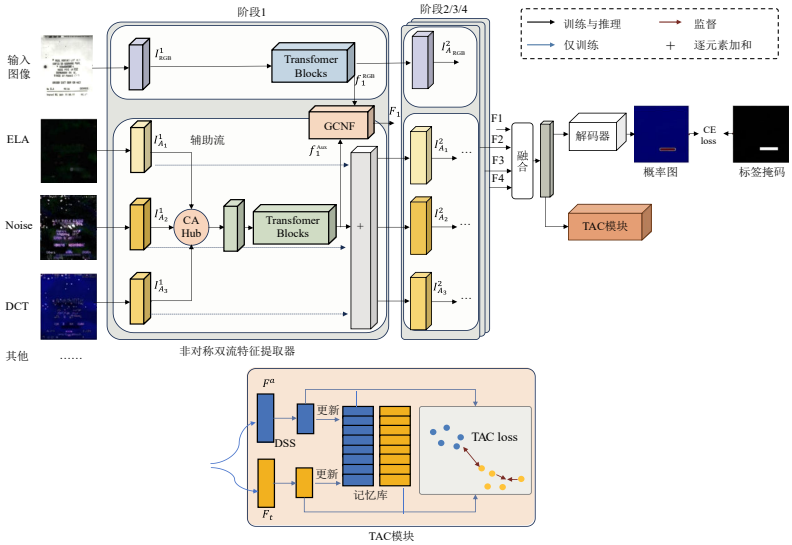


图 9 ASC-Former 模型架构

Figure 9 Architecture of the ASC-Former

除上述方法外,Ren 等人[49]提出了一种基于边缘引导的多特征融合网络 EMF-Net,用以解决文本篡改痕迹不明显且背景较为均匀情况下的篡改定位难题,如图 10 所示。该方法以 Res2Net-50 为主干,引入一系列模块来增强篡改区域的识别能力,包括通过对比学习提取篡改痕迹特征的差异语义判别模块(DSD)、融入边缘特征的边缘引导特征聚合模块(EFAM)、结合全局与局部信息的多分支注意力融合模块(MAFM),以及用于引导模型聚焦篡改区域的边缘监督模块(ESM)。在包含 12 000 张图像的文本篡改数据集 TMI-12K 上,EMF-Net 取得了优异性能,其 F1-score 和 IoU 明显高于现有基准方法(如 MVSS-Net[16]、ManTra-Net[40]),并且在 JPEG 压缩、图像缩放等常见后处理操作下仍能保持较高的定位精度,显示出良好的鲁棒性和泛化能力。

在场景文本图像中,图像语义与文本内容可能出现典型的矛盾(例如,图像背景是餐厅门口却出现“银行”字样)。因此,越来越多的研究开始关注自然场景中的文本篡改检测与定位。Wang 等人[30]将文本篡改检测引入场景文本检测任务,提出了 S3R 策略(“分割分离、回归共享”)。其在已有场景文本检测模型基础上,将像素级篡改分割任务与文本定位回归任务解耦,通过引入表示抑制损失增强对篡改纹理的敏感性。同时设计并行特征提取器,从 RGB 和频率域分别提取低层纹理和高频篡改线索并融合,以降低对大规模标注数据的依赖。作者还构建了首个字级别场景文本篡改数据集 Tampered-IC13,并在四种主流检测器上验证了 S3R 策略和频域模块的有效性。结果表明,S3R 策略平均使各模型的 mF(平均检测率)提升 2.3%~27.3%,且并行特征模块在数据减少一半的情况下仍保持检测性能。该工作首次实现了场景文本中篡改与真实文本的统一检测与区分,为从文档走向自然场景的篡改检测研究奠定了基础。

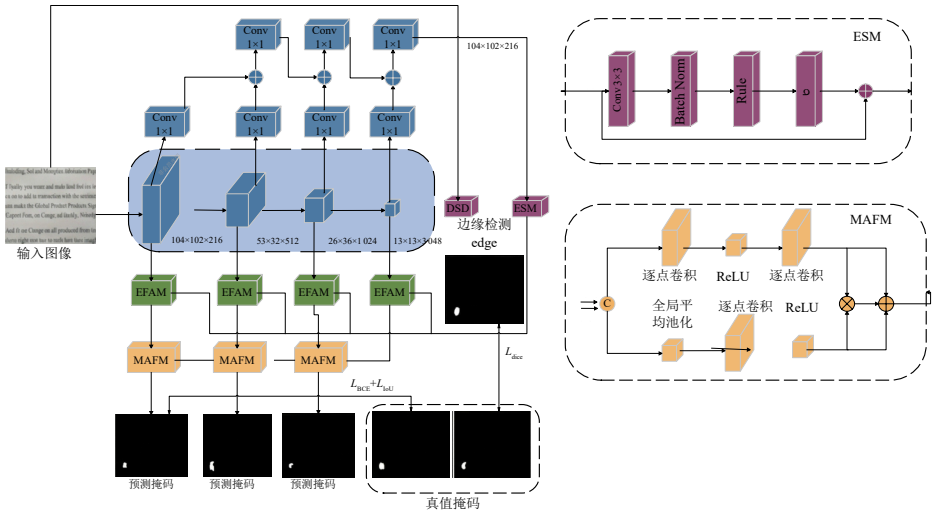


图 10　EMF-Net 模型架构

Figure 10　Architecture of EMF-Net

随着生成式 AI 驱动的场景文本编辑技术快速发展,仅针对特定篡改方式训练的检测模型在实际应用中往往面临开集泛化能力不足的问题。针对这一挑战,Qu 等人[50]提出了“生成式 AI 时代的开集场景文本篡改检测”任务,构建了 OSTF(Open-Set scene Text Forensics)基准,并设计了 Texture Jitter 预训练与 DAF(Difference-Aware Forensics)框架。OSTF 在 Tampered-IC13 的基础上,引入多种文本编辑模型生成的篡改样本,覆盖字体渲染、深度生成和扩散模型等典型编辑方式,从而在跨篡改方法、跨数据源等设置下系统评估模型的开集能力。在方法层面,作者在两阶段文本检测器基础上增加取证分支,通过 Texture Jitter 在大量真实文本图像上自动合成细粒度纹理扰动,迫使模型学习区分“真实文本—篡改文本”之间的微弱纹理差异;DAF 则借鉴无监督异常检测思想,通过学习真实文本的紧致表征并比较输入特征与该表征之间的差异来判断是否被篡改。大量实验表明,在 OSTF 和 Tampered-IC13 上,基于 Texture Jitter 与 DAF 的框架在未见篡改方法和未见场景上的 F1-score 均显著优于既有方法,零样本设置下甚至超过了部分完全监督模型,表明面向生成式编辑场景的开集建模是提升场景文本篡改分析能力的有效方向。

此外,多语言场景下的文本篡改检测也具有特殊挑战。Li 等人[51]针对包含两种语言的场景文本图像,提出了一种利用语义冲突检测篡改的方法。该方法检查图像中双语文本内容的语义一致性,比较两种语言文本描述的相似度来判定篡改;当语义相似度低于预定阈值时,即判定图像被篡改。作者还构建了首个双语场景文本篡改检测数据集 BSTID。实验结果表明,该方法在检测经二次编辑的篡改图像时表现优异,平均准确率达到 90.03%,F1-score 为 88.5%,有效验证了跨语言语义一致性检测的可行性。

总体来看,多通路融合方法的核心在于通过多模态特征融合,结合不同信息源的优势,充分利用它们在篡改检测与定位中的互补作用。文本篡改的痕迹可以出现在多个层面:视觉模态提供的像素级线索(如边缘不连续、光照不一致、局部模糊等)能够直接揭示篡改区域的细微视觉异常;频域模态通过分析压缩伪影和区块边界不连续性,暴露图像编辑和重新压缩过程中的物理痕迹;残差(噪声)模态突出相机成像噪声模式的突变,能够感知拼接操作打破原始传感器噪声一致性的区域;而 OCR 语义模态则通过金额、日期、实体属性及上下文逻辑等维度发现语义不一致或异常。在多语言场景下,基于 BSTID 等数据集的双语语义冲突检测工作,又在跨语言层面对视觉与语义模态施加了一层高阶一致性约束。通过将这些不同来源的证据进行联合建模,多流方法实际上实现了从低层物理证据到高层语义一致性的“全链路审查”。

从更细致的机理角度看,频域方法之所以在高压缩场景下更稳健,源于主流有损压缩算法(如 JPEG)本身在离散余弦变换(DCT)频域上工作的事实:图像被篡改并重新保存后,篡改区域会经历与背景不同的量化过程,形成典型的“双重压缩”模式,其在 DCT 系数上表现为具有规律性的周期性伪影。这类伪影在空间域往往与内容纹理混杂在一起而难以分辨,但在频域中更易被统计模型捕捉,因此像 DTD 这类显式建模 DCT 系数和量化表的网络[33],能够在强压缩甚至多次社交平台转发的条件下依然维持较高的定位精度。类似地,噪声残差视图会放大成像噪声模式的突变,使 EMF-Net 等方法[49]更容易感知复制、移动或

拼接操作带来的边界异常；而OCR分支则通过对文本内容进行编码，在视觉上几乎无差别的细粒度篡改场景中提供额外的高层语义判别依据。多模态多流架构通过注意力或门控机制在这些模态之间自适应分配权重：当空间域线索因压缩、模糊而退化时，频域和残差模态承担更主要的检测与定位任务；当篡改主要体现为语义逻辑冲突时，则提升语义模态的重要性，从而在更广泛的真实场景中显著弥补单流视觉模型的弱点。

在具体实现上，大多数多流架构采用“多分支特征提取+特征对齐融合”的策略：针对不同模态分别构建独立的特征提取通道，在中后期通过Transformer、交叉注意力或门控融合模块完成联合建模。DTD模型通过视觉分支和频域分支并行提取特征，再交由Swin Transformer进行统一编码，实现对压缩伪影和空间纹理的协同利用；STFL-Net借助OCR分支和RGB分支的双向跨模态注意力，使模型在保留像素级痕迹感知能力的同时，能够利用文本语义识别视觉上逼真的篡改样本；ASC-Former则采用非对称双流结构与对比学习，在真实采集场景下进一步提升了跨域鲁棒性。也有部分工作在网络早期进行“早期融合”，即将RGB、噪声残差或频域系数直接作为多通道输入，依靠浅层卷积自动学习联合表征。总体而言，多通路多模态方法在多数受控基准中、或当辅助模态与目标数据的成像/编辑链路高度匹配时，往往能在定位精度与鲁棒性上优于单模态模型；但其增益并非“必然”，在辅助域线索较弱或融合不当时也可能出现性能不升反降，因此需要在模态选择、对齐与抑噪融合机制上进行更细致的设计与验证。

3.3 基于文本语义与结构建模的方法

除了纯粹依赖视觉线索的方法，一些研究聚焦于文档内容的语义一致性和结构完整性，通过利用OCR提取的文本信息进行逻辑推理和矛盾检测。此类方法通常针对结构化文档（如收据、发票、合同等），利用语言模型、图神经网络或领域知识库来挖掘文本间的语义关系和逻辑异常。它们不局限于图像域的视觉操作痕迹，而是识别篡改行为引发的语义层级冲突，为篡改检测与定位引入了新的视角。下面介绍几种具有代表性的语义建模方法及其适用场景。

Tornés等人[52]提出了基于语言模型和领域本体的检测方法，用于识别收据中的篡改行为。该方法使用预训练法语语言模型CamemBERT，将OCR识别的收据文本转换为多种形式输入（原始文本、实体集合、实体标注文本、知识图谱三元组），以引入不同层级的结构化信息。其中，知识图谱三元组能够显式刻画收据内容中的语义关系，对检测语义不一致特别有效。作者还构建了针对收据篡改的本体结构，覆盖公司信息、支付信息、产品信息等多层次知识。在Receipts-Forgery公开数据集上进行实验时，基于三元组的表示方式在CamemBERT下获得了最高的F1-score指标，优于其他文本输入策略和传统SVM方法，也超越了纯视觉模型。这一结果表明，引入领域本体约束的语义建模对于低资源的篡改检测场景是可行且有效的，尤其适合利用先验知识来发现文本层面的逻辑矛盾。

另一种思路是将文本图像视为文本框节点构建图结构。图11展示了Joren等人[53]提出的一种基于图神经网络检测框架。作者将OCR输出的每个文本框看作图的节点，并根据空间位置关系添加边。在特征建模方面，他们首先通过变分自编码器（VAE）对节点进行预训练编码，以强调字符宽度、字体、间距等低层次特征，这有助于提升模型对潜在局部篡改痕迹的敏感性，同时尽量抑制语义信息的干扰。之后，采用多头图注意力网络（GAT）对节点特征进行关联建模，自适应调整邻近节点的权重以捕捉局部上下文。实验结果表明，在作者自建的Auto-Splice数据集（基于ICDAR 2013表格图像）上，该方法在5%篡改比例条件下F1-score达到0.904，显著优于文本专用检测方法（F1-score = 0.559）和自然图像检测方法（F1-score = 0.171）。此外，该图网络方法的平均单图处理时间仅为8.01 s，其预训练机制和图构建策略在消融实验中展现了良好的泛化性和稳定性。这些结果表明，通过图结构建模文本框之间的空间与语义关系，可以高效定位基于拼接的篡改区域，尤其适用于版式规整的文本图像篡改检测。

Guo等人[54]提出多模态篡改检测器M2F2-Det，将视觉—语言预训练模型引入图像篡改检测。该方法以CLIP为基础，通过引入少量可学习的提示与轻量适配模块，使模型在保留开放域跨模态表征能力的同时获得更强的篡改判别能力（如图12所示）。这类工作对文本图像篡改分析具有直接启发意义：其一，视觉—语言对齐提供了“文本语义—视觉呈现”一致性的强先验，可与像素级篡改痕迹形成互补，尤其有助于应对字符级微篡改、痕迹较弱但语义冲突明显的场景；其二，提示学习/适配器的范式为取证任务提供了一种低成本迁移路径，可作为现有检测/定位网络的增强分支引入而无需大规模重训主干；其三，视觉语言模型产生的相似度或注意力热图可作为候选区域先验，与分割掩码融合用于提升小目标篡改召回，并增强结果的可解释性与可审计性。

类似地，利用BLIP的图像-文本生成能力来描述篡改迹象也被探索过，但直接生成文本作为检测依据

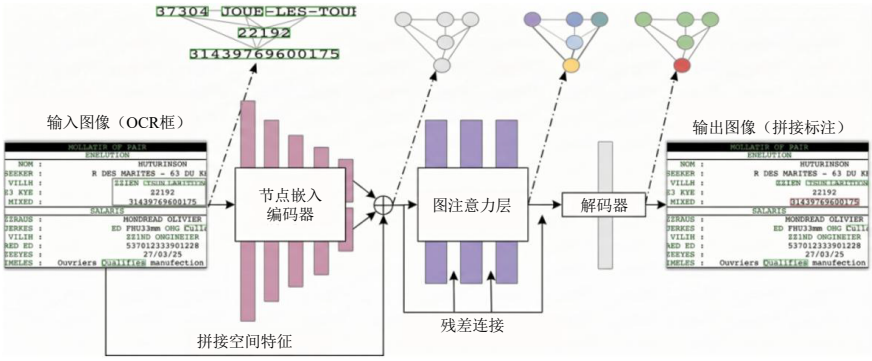


图 11　基于注意力的文档图学习模型体系结构

Figure 11　Architecture of the attention-based document graph learning model

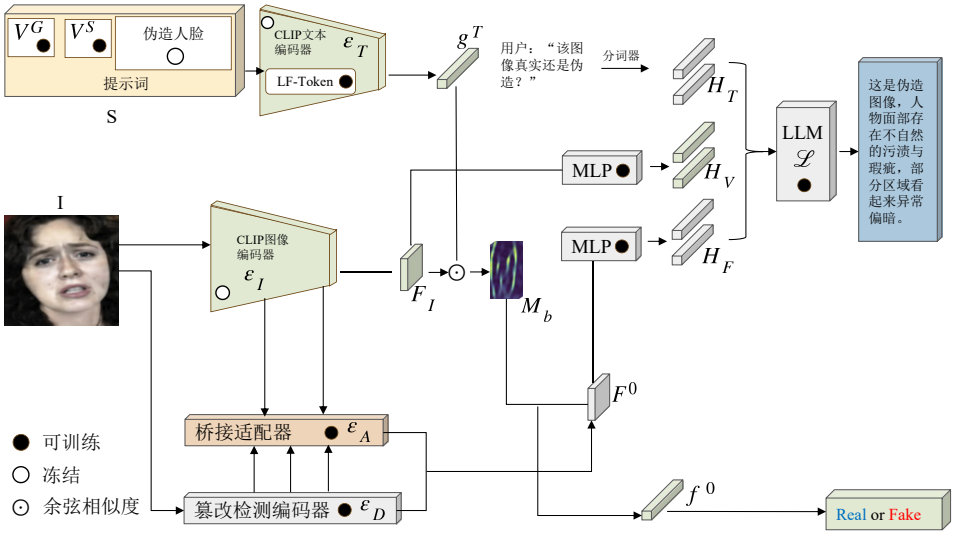


图 12　M2F2-Det 多模态篡改检测框架

Figure 12　Multimodal tampering detection framework of M2F2-Det

往往精度较低[55]。

总结来看，基于语义或结构建模的方法在检测逻辑矛盾型篡改方面具有天然优势，是对纯视觉检测的重要补充。一方面，这类方法利用了真实文档在字段依赖、版式结构和业务规则上的强约束：例如金额与大写中文的对应关系、收款方与账号信息的一致性、表格行列之间的对齐关系以及双语文本之间的语义一致性等。篡改往往只修改了其中少量字符，却会在这些关系上造成明显冲突，从而在语言模型、本体知识库或文档图结构中留下"语义异常"的信号；从"为何有效"的角度看，正是因为语义与结构约束比像素外观更加稳定，微小的文本改动才会放大为易于建模的高层不一致。另一方面，这一路线也存在明显局限：其性能高度依赖 OCR 质量和领域知识的完备程度，在扫描质量较差、版式极其复杂或跨语言场景中，文本识别错误和知识空缺都可能传递到后续推理阶段；同时，相比单流视觉模型，语义/结构建模的工程与计算复杂度更高，不适合作为极端轻量化场景中的唯一方案。因此，从跨类别对比看，纯视觉方法在"看得见"的低层物理痕迹上更敏感，多模态与频域方法在"藏不住"的统计伪影上更稳健，而语义与结构方法则在"骗不过去"的内容逻辑层面发挥关键作用，三类技术路线从不同层次共同支撑起完整的文本图像篡改检测与定位体系。

近期，部分工作也开始尝试将大型视觉语言预训练模型引入语义分析和一致性判断中，以进一步提升对复杂篡改模式的理解能力和跨模态推理能力，这为结合语言推理与视觉检测提供了新的研究方向。与此同时，无论是单流视觉网络、多通路多模态架构，还是基于文本语义和文档结构的检测框架，在设计和训练时普遍默认"测试样本分布与训练阶段基本一致"，一旦攻击者主动施加对抗扰动，或真实业务流

程中出现严重压缩、重捕获、噪声和模糊等自然降质，上述模型的判别边界仍可能变得脆弱，甚至产生系统性误判。因此，从“攻击—防御”视角系统刻画威胁模型，并设计统一的鲁棒训练机制，已经成为近年来文本图像篡改检测的重要研究方向。基于此，下一小节将专门围绕对抗攻击与防御机制展开讨论，随后在轻量化与可解释性部分进一步探讨如何在实际部署中平衡性能、开销与可信度。

3.4 对抗攻击与防御机制：模型的鲁棒训练

与自然图像分类和语义分割任务类似，文本图像篡改定位网络同样容易受到两大类扰动的影响：一类是针对检测器的对抗攻击，例如基于梯度的像素级扰动可以在几乎不改变视觉感知的前提下显著干扰篡改掩码输出；另一类是来源于真实业务流程的自然降质，如JPEG压缩、截图重采样、噪声与模糊以及社交平台传输等。这两类因素要么显式地“攻击”模型决策边界，要么隐式地削弱篡改痕迹，都会导致检测器在安全性和稳定性上的性能退化。

从“攻击—防御”的视角来看，当前针对文本篡改分析模型的攻击方式大致可以分为三类：(1)白盒对抗攻击，攻击者能够访问模型参数与梯度，典型代表包括基于PGD的像素级攻击以及针对分割任务的SegPGD攻击，可以显著降低掩码F1-score与IoU；(2)黑盒查询攻击，攻击者只能通过多次查询模型输入—输出对构造扰动，在复杂业务系统中更具现实性；(3)自然腐蚀型“攻击”，即利用高斯噪声、JPEG压缩、重采样、社交网络转发等操作破坏图像统计特性和篡改痕迹，虽然不一定出于恶意，但对于部署在真实环境中的检测器而言同样构成鲁棒性威胁。

针对上述威胁，近年来研究者提出了多种鲁棒训练与正则化策略。这一类方法的共同特点是：在不改变基础网络架构(无论其属于单流、双流还是多模态模型)的前提下，通过引入对抗样本、特征流形扰动或取证正则项，显式提升模型在对抗攻击与自然降质场景下的稳定性。下面将以潜在流形对抗训练(LMAT)、对抗取证正则化(AFR)以及自我对抗训练(SAT)为代表，综述文本图像篡改分析中的攻击－防御机制。

Shao等人[56]提出了一种基于潜在流形对抗训练(Latent Manifold Adversarial Training，LMAT)的方法，其主要面向白盒与黑盒对抗攻击，通过在特征流形上构造难例来提升分割掩码的稳健性。如图13所示，该方法在模型的潜在特征流形上施加扰动来生成对抗样本，并将其用于对抗训练，从而增强模型抵御白盒和黑盒攻击的能力。这一策略利用潜在流形上的全局信息来生成对抗样本，而不依赖直接的标签信号，避免了传统标签引导对抗训练可能导致的标签泄露和梯度掩蔽问题。在DocTamper[33]和T-SROIE[3]数据集上的实验表明，LMAT在多种攻击条件下显著提升了篡改定位模型的鲁棒性，尤其在白盒和黑盒攻击下表现出色，整体上超过了传统对抗训练方法(如PGD-AT[57]、SegPGD-AT[58]、Trades-AT[59])。值得注意的是，LMAT在不明显降低模型对干净样本检测与定位精度的前提下，大幅提高了模型定位篡改区域的能力，展现出良好的实用潜力。

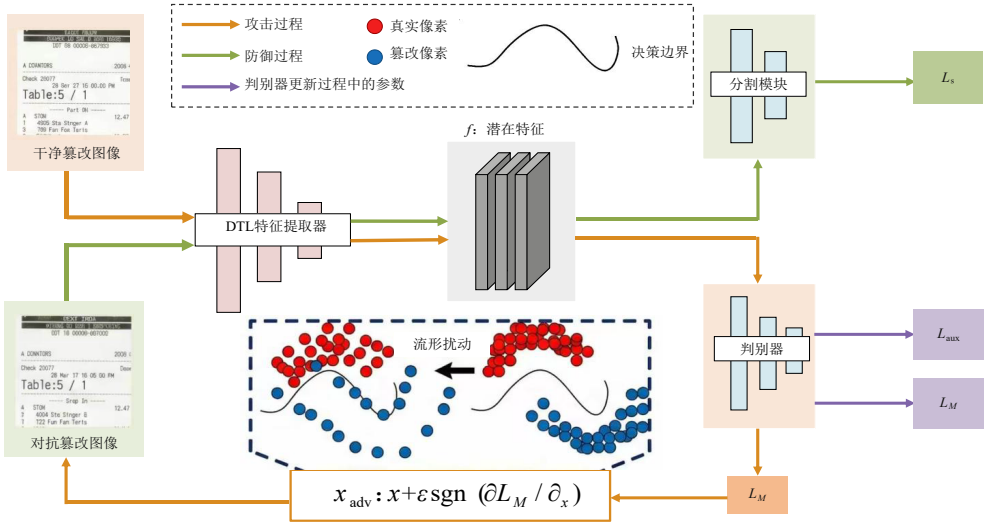


图13 基于潜在流形的对抗训练整体框架

Figure 13 Overall framework of latent manifold adversarial training

该作者进一步提出了一种基于最小-最大优化的对抗取证正则化方法AFR(Adversarial Forensic Regularization)[60],其更侧重于自然降质(如噪声、压缩)情况下的鲁棒性,在兼顾对抗攻击防御的同时,显式建模篡改区域与真实区域之间的取证相关性。AFR利用互信息度量篡改区域与真实区域之间的取证相关性,并通过最小化该相关性来减少二者共享的信息,从而提升模型在图像自然损坏下的鲁棒性。此外,为平衡模型在干净图像和受损图像上的预测分布,作者设计了分布间散度正则化(DDR)项,进一步增强模型的鲁棒性与泛化能力。在T-SROIE[3]数据集上的实验中,与传统对抗训练方法(如PGD-AT[57]、LMAT[56]等)相比,AFR在处理各种自然干扰(高斯噪声、JPEG压缩、社交网络传输等)时显示出明显优势,其中针对社交网络传输造成的篡改损坏,F1-score提升了46.9%。同时,AFR在应对对抗性攻击时也表现出色,整体鲁棒性优于现有对抗训练方法。

此外,Zhuo等人[61]提出自我对抗训练(Self-Adversarial Training,SAT),从训练策略层面增强模型在弱痕迹与复杂后处理条件下的稳定性。其核心思路是利用模型自身的响应(如注意力/显著性或预测不确定性)自动挖掘“更难”的区域,并在这些区域内生成对抗扰动用于迭代训练,从而在样本有限的情况下持续强化判别边界。虽然SAT主要在通用图像篡改场景中展开,但其“以模型自身为导向的难例生成”对文本图像同样具有启发意义:例如可将扰动约束在OCR文本框或疑似篡改区域内,模拟字符级篡改在压缩、重采样、截图传播后导致的痕迹衰减;同时配合频域/残差线索的约束,使模型在对抗扰动与自然降质叠加时仍能输出更稳定的篡改掩码。

从整体策略上看,LMAT、AFR与SAT分别对应三类互补方向:LMAT侧重面向显式对抗扰动的鲁棒优化,通过在特征流形上构造难例提升模型在白盒/迁移攻击下的稳定性;AFR更强调真实业务链路中的自然降质,通过取证相关约束与分布一致性正则提升在压缩、噪声、重采样与社交传播等条件下的可靠性;SAT则在样本有限、弱痕迹与多种后处理组合更常见的场景中,提供了一种以“自生成难例”增强训练信号的思路。三者共同指向一个趋势:鲁棒性提升不再依赖单一攻击假设,而是更面向真实链路的复合扰动与跨域泛化需求。

为便于读者从“攻击—防御”视角整体把握不同方法的适用场景与优缺点,本文在表3中进一步总结了主流对抗训练与鲁棒正则化方法的特性对比,包括其主要应对的威胁模型(对抗攻击/自然降质)、是否依赖标签或梯度信息、对基础检测器结构的侵入程度以及在典型数据集上的鲁棒性能变化趋势。可以看到,潜在流形增强类方法(如LMAT)主要面向显式对抗攻击,自然腐蚀鲁棒化方法(如AFR)更适合部署在存在复杂传输链路的真实系统中,而自我对抗训练则在数据稀缺条件下提供了一种兼顾鲁棒性与可实现性的折中方案。整体而言,将这些防御策略与前述3.1~3.3节介绍的不同检测网络相结合,有望构建起从“架构设计—模态融合—语义建模—鲁棒训练”多层次协同的文本图像篡改检测与定位体系。

表3　典型“攻击—防御”方法对比

Table 3　Comparison of representative attack-defense methods

方法名称	主要防御对象	依赖信息	对基础检测网络的改动程度	对鲁棒性的影响	对干净样本测试集性能的影响	典型适用场景与特点
标准监督训练(Standard)	无特定防御对象,默认同分布	仅需标注数据	无结构改动	对攻击和降质整体较脆弱	干净样本性能通常最高	无明显攻击、质量较好的离线分析
像素空间对抗训练(PGD-AT等)	白盒/近白盒对抗攻击	需访问梯度或损失	训练中加入对抗样本与损失	对对应攻击鲁棒性明显提升,对自然降质有限	一般略有精度下降	安全敏感、模型结构已知的场景
自然腐蚀数据增强(JPEG等)	压缩、噪声、重采样等自然降质	不需攻击知识,仅需链路先验	仅在数据层面增加失真样本	对相应降质鲁棒性有一定提升	对干净样本性能影响较小	社交平台、即时通信等复杂传输链路
潜在流形对抗训练(LMAT)	白盒/黑盒对抗攻击,兼顾部分降质	需特征空间梯度、篡改标签	主干结构不变,在特征流形上加入对抗扰动与正则	对多种对抗攻击鲁棒性显著增强,对部分降质也有提升	基本保持干净样本性能	既重视安全又要求高精度定位的系统
对抗取证正则化(AFR)	自然降质为主,兼顾部分攻击	不依赖具体攻击算法,依赖取证统计信息	主干基本不变,增加取证相关正则项	多种降质下预测更稳定,F1-score提升明显	合理设置权重时可基本不降甚至略升	真实业务链路复杂、压缩重采样频繁的部署环境
自我对抗训练(SAT)及相关方法	弱痕迹篡改、多种后处理组合,有限样本场景	利用模型自身注意力/预测生成“自对抗”样本	主干不变,增加自对抗样本生成与损失	提升弱篡改、多后处理和跨数据集鲁棒性	一般对干净样本性能影响较小	真实篡改样本稀缺、标注成本高的应用环境

3.5 轻量化检测方法与模型可解释性实践

需要指出的是，上述对抗训练与鲁棒正则化方法，在提升安全性与稳定性的同时也不可避免地带来了额外的训练与推理开销，这在资源受限的移动端与边缘侧部署中尤为突出。此外，对抗鲁棒模型的决策边界往往更加复杂，进一步加剧了模型内部机理难以直接解释的问题。因此，在鲁棒性之外，如何在保证检测性能的前提下兼顾轻量化与可解释性，成为推动文本图像篡改检测走向实际落地的另一条关键技术路线。本节将围绕这两个方面对现有工作进行综述与分析。

从应用需求出发，模型的轻量化与可解释性已经成为文本篡改检测领域亟需关注的新方向。一方面，实际场景下移动设备部署和实时处理要求检测模型具备较低的计算复杂度和延迟；另一方面，高可信场景（如司法鉴定、金融风控）需要模型给出透明可信的判别依据。正因如此，研究者开始尝试在保持篡改检测性能的同时，通过模型压缩、知识蒸馏等手段打造轻量高效的检测网络，并探索结合人类可理解的解释输出以提高模型决策的透明度和用户信任度。在本节中，我们将介绍文本图像篡改检测中轻量化检测方法的最新进展，以及结合模型可解释性的实践方案。

在轻量化部署方面，尽管目前专门针对文本图像篡改的轻量化模型较少，但邻近领域（如通用图像拼接定位）的研究提供了宝贵思路。Zhao 等人[62]提出的 Mobile-PspNet 以 MobileNetV2 风格的轻量编码器为骨干，并结合金字塔池化进行多尺度上下文聚合，在较低计算开销下仍能保持较好的篡改区域分割效果，并在 CASIA[63]、COLUMB[64]等数据集上验证了其有效性。对文本图像篡改任务而言，这类“轻量骨干+多尺度上下文”的组合具有直接参考价值：一方面，轻量编码器有助于降低移动端/边缘端的延迟；另一方面，多尺度聚合可覆盖字符级微篡改、跨字符/跨行的文本块篡改等尺度变化带来的挑战。进一步地，若结合文本区域先验（如 OCR 文本框、版面分析结果）进行注意力引导，或叠加轻量的频域/残差辅助分支，往往可以在参数增长可控的前提下增强对弱痕迹与强压缩场景的适应性，从而更贴近实际部署需求。

这启示我们，在文本图像篡改检测与定位领域，同样可以通过缩减模型深度来获得轻量高效的模型。此外，亦可借鉴极浅网络[65]、知识蒸馏和模型剪枝思想，可参考深度篡改检测领域已有的模型压缩方法来提升文本篡改检测模型的推理效率。

而随着篡改检测模型复杂度不断提高，可解释性成为关键需求。Qu 等人[66]提出了一种可解释的文本篡改检测方法，旨在利用多模态大型模型来检测被篡改的文本，并生成自然语言描述解释检测结果（如图 14 所示）。在该任务中，作者提出了一种基于 GPT-4 的文本生成方案，用于描述被篡改文本在视觉和语言层面的异常。为减少模型误判，作者创新性地设计了“融合掩码提示”（fused mask prompt）策略，通过像素加权将篡改区域掩码与原图融合，引导模型更准确地定位篡改区域并生成异常描述。此外，他们提出的篡改文本侦测器（Tampered Text Detective，TTD）模型通过引入定位提示，使得大模型能够更有效地聚焦可疑篡改区域，提升模型的细粒度感知能力并减少误检。广泛的实验表明，TTD 模型在自建的 ETTD 数据集和公共 T-IC13 数据集[30]上均取得了显著成绩，超越了现有基准模型，表现出优秀的领域内和跨领域泛化能力。这表明该方法在可解释的文本篡改检测任务中具有显著优势，推动了该领域进一步的发展。

图 14 TTD 模型检测结果

Figure 14 TTD model detection results

类似地，Guo 等人[54]提出的 M2F2-Det 模型将 CLIP 与大语言模型结合，在实现高精度篡改检测的同时生成细粒度的文本解释，为多模态可解释取证提供了有益探索。这些最新工作表明，将轻量化网络设计与模型可解释性方法有机结合，一方面可以在移动端或业务系统中降低部署成本、提升推理效率，另一方面也有助于增强用户和监管者对自动化判别结果的信任，为构建可落地的文本图像篡改检测系统奠定基础。总体而言，轻量化与可解释性为前文介绍的单流视觉、多模态融合以及语义与结构建模三类方法提供了一个新的“工程维度”：不仅要关注检测与定位精度和鲁棒性，还需要在模型规模、延迟和可审计性之间做出合理折中。基于这一考虑，在随后小节的性能与复杂度对比中，我们将综合比较代表性方法在定

位性能和模型开销上的差异,为不同应用场景下的方案选择提供参考。

3.6 代表性方法性能与复杂度对比

为了更直观地比较不同技术路线在真实场景、跨域泛化与工程开销上的差异,本文在前述综述基础上选取若干典型模型,并从“技术路线概览—真实基准对比—跨域测试—复杂度权衡”四个层面构建对比:表4归纳各路线的核心设计与适用场景;表5与表6在真实手工篡改基准RTM上统一报告像素级定位性能,用于检验各路线在弱痕迹、小区域场景下的能力边界;表7进一步给出T-SROIE数据集上的跨域泛化结果,以观察模型在不同数据分布下的迁移表现;表8汇总部分方法的参数规模与推理效率,为不同算力条件下的方案选型提供工程参考。

由表4可以看出,单流视觉路线整体结构较简洁、训练与部署门槛较低,但在“弱痕迹+小区域+非规则边界”的真实篡改条件下,单一视觉通路往往更容易面临漏检或误报的权衡瓶颈。值得注意的是,单流并不等价于“性能一定弱”:近年来基于Transformer的通用分割框架由于具备更强的全局关系建模能力,在一定程度上弥补了局部纹理线索不足的问题,但其证据仍主要来自RGB域,面对取证场景所要求的“低误报+可信定位”仍存在上限。多模态方法通过引入频域、残差或误差域等互补线索,有潜力在误报控制与定位可信度上取得优势,但其收益高度依赖辅助模态与目标数据成像/压缩机制的匹配程度,以及融合策略对噪声线索的抑制能力。基于语义/结构一致性的路线则在内容逻辑层面提供额外约束,但对OCR与版面解析质量较敏感。基于此,下面以RTM为统一基准对不同路线进行定量对比与分析。

此外,为了定量展示各方法在真实文本篡改场景下的定位效果,本文选用近期提出的RTM数据集作为统一评测基准。相比大量合成篡改数据,RTM的篡改均由人工真实操作完成,且包含更加丰富的篡改类型;更关键的是其篡改区域通常面积更小、形状更不规则、边界更弱,使得篡改痕迹更隐蔽、更接近真实场景中的取证难度。因此,在该数据集上多数方法难以取得理想性能,结果普遍偏低但更具区分度,也更能反映不同技术路线在“弱痕迹、小区域”条件下的真实能力。

表4　代表性方法的技术路线与优缺点概览

Table 4　Overview of technical routes, advantages, and limitations of representative methods

方法	技术路线	复杂度 / 部署	主要优点	主要局限
Dense-FCN[19]	单流视觉(RGB)	中等:编码-解码型FCN,参数量适中,训练收敛稳定	结构相对简单;对JPEG压缩、缩放、高斯噪声等常见后处理具有较好鲁棒性,可作为通用基线	主要依赖像素级纹理线索,对强后处理、复杂版式和语义层面篡改的感知有限
MFAN[38]	单流视觉+多级特征注意	中等:以ResNet-50为主干,叠加多级通道与空间注意模块,可在单卡上完成高分辨率证书图像的训练与推理	针对证书等复杂版式文档进行定制化建模,在文档类场景具有较好的泛化能力	主要面向文档场景优化,对自然场景文本或多语种数据的适应性仍需进一步验证;仍完全工作在视觉域,对金额、条款等语义层面篡改的直接建模能力有限
TIFDM[35]	单流视觉+多尺度注意	中等偏高:引入FTFN、LHSE、TFAM等多模块,显存和推理开销略高于普通U-Net	对JPEG压缩、重采样等弱痕迹场景有较强定位能力,对小目标和模糊边界的定位效果好;在多数据集上表现稳定	仍主要依赖视觉模态,对语义层矛盾和复杂版式利用有限;网络结构相对复杂,不利于极端轻量化部署
DTD[33]	多模态双流+Transformer:RGB视觉分支+DCT频域分支+Swin Transformer	高:Swin Transformer主干,多分支与课程学习训练,适合离线/服务器端部署	频域+视觉互补,对压缩、模糊等退化场景鲁棒;跨域测试上优势明显	模型参数量和计算量大,实现和调参成本高,对算力和数据规模要求较高;对高分辨率长文档的端到端处理成本较高,不利于移动端实时部署
双流[29](SIEN+CFEN)	多模态双流:空间外观(SIEN)+像素相关性/残差(CFEN)	中等:卷积双分支+判别网络,训练与推理仍可在单卡完成	将空间纹理与像素相关性联合建模,对缩放、复制粘贴等操作的物理痕迹较敏感;区域级定位精度较高	依赖人工设计的残差滤波器,迁移到新型篡改或不同成像设备时需要重新调参;对语义信息几乎未利用

续表

方法	技术路线	复杂度 / 部署	主要优点	主要局限
ASC-Former[5]	非对称多流：RGB主干流+频域/SRM/ELA多模态辅助流+Transformer聚合	高：包含多模态分支、CA Hub与GCNF等注意力模块，并引入对比学习TAC，训练与推理成本较高	可灵活接入多种取证模态，并通过门控注意力自适应选择最有判别力的线索；在真实采集的中文场景中表现优越，跨域泛化能力好	结构复杂、参数量大，对显存和算力要求高；整体依赖高质量的多模态预处理（频域、SRM、ELA等），工程集成成本较高
图模型（GAT）	语义/结构一致性：基于文本框构图+多头图注意力	中等：节点数与文本框数量相关，GAT推理成本中等；适合离线审核或批处理	显式建模文本框之间的版面结构与邻域关系，对表格、证书等版式规整文档中的拼接篡改特别敏感；对训练数据需求相对较小	严重依赖OCR与版面分析质量；对非结构化场景文本或复杂自然背景不够适应；实时性有限
双语语义一致性检测	语义一致性：跨语言文本相似度度量	低：主要基于文本特征与相似度计算，模型较轻，可在普通服务器甚至高端终端上运行	直接利用双语语义冲突进行判别，对那些视觉上自然但语义不一致的篡改（如招牌内容不符）有明显优势	对OCR质量和机器翻译/语义编码准确度高度敏感；仅适用于存在多语言冗余的场景，对单语文档或结构化票据支持有限
LMAT[56]/AFR[60]	鲁棒性增强：潜在流形对抗训练（LMAT）+对抗取证正则化（AFR）	训练阶段复杂度高：需生成潜在空间对抗样本并加入最小-最大优化；推理阶段与原模型接近，可沿用原部署	在不显著损失干净样本精度的前提下，大幅提升模型对对抗攻击和自然降质的鲁棒性；可作为已有检测器的“增强插件”	训练成本高，对算法工程实现和超参数敏感；理论与实现门槛较高，目前主要在研究原型中验证，距离大规模工程落地仍有距离
Mobile-PspNet[62]	轻量化单流：MobileNetV2主干+金字塔池化多尺度融合	低：基于MobileNetV2的轻量结构，适合移动端与嵌入式部署，可实现接近实时的推理	在保证较高定位精度的同时显著降低模型复杂度，为资源受限环境下的篡改定位提供可行方案	针对的是通用图像拼接任务，未专门考虑字符级微篡改与语义一致性；直接迁移到复杂文本图像时仍需适配和再训练
TTD[66]	可解释多模态：检测网络+多模态大模型+融合掩码提示	高：需要篡改定位网络+大模型推理，计算和显存开销较大，更适合离线分析或人工复核场景	能在给出像素级篡改定位的同时生成人类可读的解释，显著提升模型透明度和可审计性，适用于司法取证、金融风控等高可信场景	对计算资源要求高；目前解释质量仍依赖大模型能力和掩码提示设计，在极端复杂版式和多语言场景下的稳定性有待进一步验证

为保证对比的代表性与可复现性，本文从三条路线选取基线共同评测：（1）通用语义分割框架，其中CNN-based包括UperNet、DeepLabV3+与HRNet-OCR、Transformer-based包括SegFormer、MaskFormer与Mask2Former；（2）通用图像篡改定位基线；（3）面向文本篡改的专用方法（第3章所述代表性单流与多流模型）。这种设置既能横向对照“语义分割式定位”与“取证式定位”的差异，也便于观察多模态线索融合在真实弱痕迹场景下的增益。所有对比方法均在RTM训练集上训练。

从RTM测试集统一评测的定量结果来看（表5、表6），在更贴近真实篡改的场景下，各方法的像素级定位上限仍然有限：在包含真实样本的“全图”口径下，当前最优方法的IoU仅为19.71%，像素级F1-score最高为32.93%。这一结果说明，在真实场景更关注的“低误报+可信定位证据”要求下，现有模型仍难以稳定地在细粒度层面给出准确掩码，整体距离实际应用仍有明显差距。下面结合表5、表6，进一步分析不同技术路线的优势与局限。

表5给出了RTM测试集上不同篡改类型的IoU，并分别报告“总篡改”与“全图（All）”两种统计口径。首先，从篡改类型看，拼接与修复往往引入来自外部内容或背景再生成带来的局部统计不一致，因此整体IoU相对更高；相比之下，复制移动与覆盖多来自同图区域，局部纹理/噪声与周围更一致，而插入往往落在字符笔画级的极细粒度区域，可利用的边界与纹理差异更弱，是RTM上普遍的主要难点。

其次，从“总篡改”到“全图”的变化可以侧面反映误报控制能力：部分方法在“总篡改”上看似可观，但在包含真实样本的“全图”口径下出现明显回落，说明其对真实样本更容易产生误报（例如Mask2Former从17.18%下降到12.35%，Liang等人[14]从8.39%下降到4.63%）；相对地，SegFormer（17.66%下降到15.72%）与ASC-Former（21.57%下降到19.71%）的降

表5　代表性方法在RTM测试集上的像素级IoU表现　　单位:%

Table 5　Pixel-level IoU of representative methods on the RTM test set　　unit:%

类别	代表性方法	来源	不同篡改方式						总篡改	全图
			复制后移动	拼接	插入	覆盖	修复	编辑		
语义分割方法	UperNet[67]	ECCV 2018	6.67	11.09	4.66	11.53	9.14	10.99	8.92	8.26
	DeepLabV3+[68]	ECCV 2018	7.32	12.75	2.57	9.16	10.74	13.69	9.26	8.56
	HRNet-OCR[69]	TPAMI 2021	6.06	13.16	0.74	6.18	7.75	11.35	7.61	6.81
	SegFormer[70]	NeurIPS 2021	14.01	25.51	15.77	13.22	27.60	17.53	17.66	15.72
	MaskFormer[71]	NeurIPS 2021	18.27	22.11	16.95	11.11	17.28	18.58	17.23	13.72
	Mask2Former[72]	CVPR 2022	15.43	20.37	12.96	14.49	23.31	16.59	17.18	12.35
图像篡改基线	RRU-Net[15]	CVPRW 2019	2.39	6.20	3.39	3.54	7.91	4.07	4.08	3.73
	PSCC-Net[73]	TCSVT 2022	4.52	7.49	1.31	4.73	6.43	3.34	4.93	3.31
	MVSS-Net++[45]	TPAMI 2023	8.75	12.06	4.65	7.18	5.18	12.24	8.43	5.11
	CAT-Net v2[74]	IJCV 2022	11.7	17.28	6.62	10.97	15.48	8.09	12.64	11.30
第三章方法	Liang等人[14]	TrustCom 2022	6.16	15.55	3.10	8.30	10.52	12.22	8.39	4.63
	TIFDM[35]	TCE 2024	5.13	15.45	4.48	0.17	0.16	6.91	5.02	3.65
	DTD[33]	CVPR 2023	7.65	12.93	6.26	8.79	10.91	6.76	8.98	6.51
	ASC-Former[5]	Pattern Recognit 2025	18.57	32.79	18.89	16.06	27.63	19.35	21.57	19.71

幅更小,体现出在保持检出的同时更好的误报抑制。

从方法类别看,经典CNN分割模型整体IoU偏低(全图约6.81%~8.56%),主要受限于对小区域篡改的召回能力;而基于Transformer的分割框架显著更强,其中SegFormer在全图IoU = 15.72%、总篡改IoU = 17.66%,反映出全局注意力对版面/字符间不一致关系建模的优势。值得注意的是,第3章专用方法在RTM上呈现出更强的“路线差异”:两种单流方法Liang与TIFDM在全图IoU上分别仅为4.63%与3.65%,说明面向特定干扰(如混合、压缩/重采样)设计的单通路特征增强策略在更真实的弱痕迹分布下可能出现明显域偏移;DTD虽然引入DCT频域分支,但全图IoU仅为6.51%,未能超过强单流基线SegFormer,提示“多模态”并非必然带来收益,其效果高度依赖辅助模态是否与RTM的成像/编辑链路高度相关以及融合时能否抑制噪声线索。相对地,ASC-Former通过多模态辅助流与门控聚合机制在各类篡改类型上均取得最优,全图IoU提升至19.71%,表明在RTM这类真实手工篡改条件下,动态选择并融合互补线索仍是提升定位可靠性的有效方向。

观察表6可以发现,在RTM这种“篡改像素占比极低、篡改区域细碎且边界弱”的设置下,不同方法往往呈现明显的Precision-Recall权衡:经典CNN分割模型(UperNet/DeepLabV3+/HRNet-OCR)普遍表现为“精度相对较高但召回偏低”,说明其更容易漏检细小篡改区域;相反,部分方法会通过更激进的预测换取更高召回,但带来严重误报,例如PSCC-Net召回率达到30.28%,但精度仅3.59%,像素级F1分数仅6.41%。这也解释了为何某些方法在图像级F1分数上可能较高(如MaskFormer为68.83%、PSCC-Net为68.71%),但像素级定位并不理想:图像级任务更关注“是否存在篡改”,而像素级定位更强调“证据是否可信且边界是否准确”,二者并不总是正相关。

聚焦第3章方法可以看到,单流模型方法在RTM上的不足主要体现在误报与漏报两端:Liang等人呈现“召回不低但精度极低”(Recall = 24.99%, Precision = 5.37%),说明其更倾向于在弱纹理/弱边界条件下过度响应,从而造成大面积误报;TIFDM则同时表现为精度与召回偏低(Precision = 6.52%, Recall = 7.67%),反映其在RTM分布下更偏保守,容易漏检细碎篡改。DTD在精度与召回上相对均衡(11.94%/12.52%),但总体F1分数仍仅12.22%,表明仅引入单一频域线索并不足以稳定提升真实手工篡改下的像素级证据质量。相比之下,ASC-Former在保持相近召回(24.44%)的同时将精度显著提升到50.44%,从而将像素级F1分数提升至32.93%,更符合真实取证场景对“低误报+可审计定位证据”的需求。

从“单流—多流”的角度看,RTM结果揭示出一个更细致的结论:单流并非必然弱,但其上限高度依赖骨干表征与全局建模能力。以SegFormer为代表的单流Transformer分割框架在RTM上已具备较强竞争力(F1-score = 27.17%),说明仅凭RGB也能学习到部分版面/文本不一致线索;但其性能仍明显低于ASC-Former(32.93%),表明当篡改痕迹进一步弱化时,仅依赖单域线索仍难以兼顾召回与误报控制。另一方面,多流方法也并非“天然更强”:DTD在RTM上不如

表6 代表性方法在RTM测试集上的像素级精度、召回率、F1分数和图像级F1分数 单位:%

Table 6 Pixel-level precision, recall, F1-score, and image-level F1-score of representative methods on the RTM test set unit:%

类别	代表性方法	来源	像素级			图像级
			精度	召回率	F1分数	F1分数
语义分割方法	UperNet[67]	ECCV 2018	32.49	9.98	15.27	49.11
	DeepLabV3+[68]	ECCV 2018	32.24	10.43	15.76	52.87
	HRNet-OCR[69]	TPAMI 2021	24.15	8.67	12.76	47.00
	SegFormer[70]	NeurIPS 2021	38.45	21.01	27.17	61.51
	MaskFormer[71]	NeurIPS 2021	26.00	22.50	24.13	68.83
	Mask2Former[72]	CVPR 2022	19.10	25.89	21.99	64.02
图像篡改基线	RRU-Net[15]	CVPRW 2019	15.20	4.72	7.20	40.97
	PSCC-Net[73]	TCSVT 2022	3.59	30.28	6.41	68.71
	MVSS-Net++[45]	TPAMI 2023	7.34	14.41	9.73	54.87
	CAT-Net v2[74]	IJCV 2022	30.18	15.30	20.31	54.82
第三章方法	Liang等人[14]	TrustCom 2022	5.37	24.99	8.84	59.64
	TIFDM[35]	TCE 2024	6.52	7.67	7.05	56.81
	DTD[33]	CVPR 2023	11.94	12.52	12.22	53.76
	ASC-Former[5]	Pattern Recognit 2025	50.44	24.44	32.93	63.31

SegFormer,恰恰说明多模态收益取决于辅助域是否与目标数据高度相关,以及融合策略是否能在复杂噪声下稳定筛出有效证据。因此,面向真实部署,更稳健的路径往往是“强单流表征+可控的多模态补充”,并通过门控/注意力等机制将多流增益转化为可解释的误报抑制与证据增强。

表7 代表性方法在T-SROIE测试集上的泛化效果 单位:%

Table 7 Generalization performance of representative methods on the T-SROIE test set unit:%

方法	SegFormer[70]	MaskFormer[71]	CAT-Net v2[74]	DTD[33]	ASC-Former[5]
IoU	64.6	71.89	55.03	55.41	72.06

为进一步观察模型在不同数据分布与标注形态下的迁移表现,本文选取表5中覆盖不同路线且具有代表性的若干方法进行跨域测试(结果如表7所示):所有方法均在RTM训练集上训练,并在T-SROIE测试集上评估。可以看到,单流Transformer分割模型在该设置下已能获得较高IoU(SegFormer为64.60%、MaskFormer为71.89%),说明当目标数据的篡改形态更规则、区域更接近水平矩形框且痕迹更显著时,通用分割框架具备较强的适配能力。ASC-Former进一步达到72.06%,但与MaskFormer的差距已非常有限,这表明在相对“更容易”的合成票据基准上,强分割基线往往已接近性能上限,多模态带来的增益会被压缩。与之形成对照的是,RTM基准上各方法IoU普遍不足20%,更能体现真实弱痕迹场景的困难性与技术路线差异。因此,仅在合成基准上取得高分并不能充分代表对真实手工篡改的有效定位能力;后续研究仍需在更真实的数据分布与更严格的误报控制条件下持续提升跨域鲁棒性。

从表8可见,推理效率与参数规模并非一一对应,而是由“参数量+结构复杂度(多分支、多尺度解码、跨模态融合、预处理链路等)”共同决定。总体上,移动端轻量模型(如Mobile-PspNet)具备“非常快”的推理效率;小规模单流卷积网络或以密集连接/空洞卷积为主的结构(如Dense-FCN)多处于“快”或“中等”区间。引入双流/多分支与多尺度融合(如MVSS-Net、STFL-Net、EMF-Net),或采用Transformer/注意力作为核心建模机制(如ASC-Former、DTD)时,计算与显存开销上升,推理速度多归为“中等”;当主干达到百M级别或采用更大主干(如FFDN、CAT-Net、MVSS-Net)时,推理速度通常下降至“慢”。值得注意的是,结合表5~表6的性能结果可以发现,“更大模型/更复杂结构”并不必然带来真实RTM基准上的更好定位效果:在真实弱痕迹场景中,模态选择是否匹配、融合是否能抑制噪声以及误报控制策略往往比单纯堆叠规模更关键。该表旨在提供工程量级的对比视角,以辅助不同应用场景下对精度、速度与可审计性的综合权衡。

综合表4~表8可以看到,当前文本图像篡改分析方法在“真实弱痕迹定位性能—跨域泛化—工程开销—可解释性”之间形成了明确的多维权衡:强单流

表8　代表性方法的参数规模与推理速度对比

Table 8　Comparison of parameter scale and inference speed of representative methods

类别	代表性方法	核心复杂度要素	参数量/M	推理速度
基线方法	ManTra-Net[40]	CNN基线(无Transformer/无多分支)	4.0	快
	MVSS-Net[45]	双分支ResNet+多尺度融合	143.0	慢
	CAT-Net[74]	高分辨率主干(HRNet)+复杂解码	114.0	慢
第三章代表性方法	TIFDM[35]	EfficientNet-B3+注意力模块	≈20.0*	中等
	MFAN[38]	注意力融合+多尺度	≈30.0*	中等
	Dense-FCN[19]	密集连接+空洞卷积	≈12.3*	快
	Xu等[29]	双分支特征增强(SIEN/CFEN)	≈10.0*	快
	CTP-Net[42]	双流CNN(CTS+ITS)	≈10.0*	快
	DTD[33]	Transformer(Swin)+模块增强	66.0	中等
	FFDN[46]	大模型(ConvNeXtV2-B)+多尺度	140.0	慢
	STFL-Net[47]	双流+OCR引导	≈55.0*	中等
	EMF-Net[49]	边缘引导+特征融合(多分支)	≈28.0*	中等
	ASC-Former[5]	Transformer(MiT)+注意力融合	≈27.0*	中等
	Joren[53]	图推理(GCN)	—	非常慢
	TTD[66]	大模型(VLM-7B)	≈7 000*	非常慢
	Mobile-PspNet[62]	轻量主干(Mobile)+金字塔池化	0.66	非常快

注:1.参数量带“*”为估算值:以论文明确给出的主干网络为基准,参考该主干的常见公开参数规模,并结合是否存在多分支、多尺度解码、注意力融合等模块带来的额外参数增量进行近似估计;2.为避免不同论文硬件/输入分辨率差异带来的不可比性,本文将推理效率按模型规模与结构复杂度划分为五档,具体为:非常快,移动端/轻量模型(参数量<1 M),或明确面向实时部署的轻量架构;快:小模型(约≤15 M),且结构较轻(单流CNN为主),无重型Transformer主干/无明显多分支开销;中等:中等规模(约15~70 M),或包含较多注意力模块、多分支结构、复杂融合等带来额外计算开销;慢:大模型(约>70 M),或采用图结构推理、生成式推理等计算链路较长的方法;非常慢:秒级推理或B级参数量大模型(如VLM-7B及以上)。

Transformer分割框架在RTM上已具备较强竞争力,但像素级性能上限仍偏低;多模态/多流方法在误报控制与证据可信度上更具潜力,但其收益依赖于辅助模态与真实成像链路的匹配以及融合策略的抑噪能力,同时引入额外实现与部署成本;语义/结构一致性与大模型可解释路线为高可信应用提供补充证据,但对OCR质量与推理开销更敏感。面向实际落地,如何在强单流表征基础上以“可控的多模态补充”提升真实场景下的定位可靠性,并在误报管理、算力预算与可审计需求之间实现可解释的折中,将是后续研究与工程部署需要重点解决的问题。

4　总结与展望

随着深度学习与多模态技术的持续演进,文本图像篡改分析在多个任务指标上已取得显著进展,然而距离在现实复杂环境中的广泛应用仍存在显著差距。本章围绕当前技术发展的核心挑战、典型方法的局限性、部署实践的实际障碍以及法律伦理层面的外部推动,展开深入讨论,并指出未来研究的潜在方向。

4.1　技术挑战与研究瓶颈

(1)真实场景泛化能力不足:当前大多数检测与定位模型依赖合成数据集进行训练,虽便于大规模建模,但在跨领域、跨模态场景中仍面临严重的性能退化问题。例如,从票据转向身份证、合同或自然场景图像时,检测与定位精度显著下降。这主要源于模型对特定字体、排版、背景的过拟合,难以处理真实世界中的扫描失真、图像压缩和非标准排版。

(2)对抗性篡改与鲁棒性威胁:随着扩散模型和GAN等生成技术的发展[75],攻击者可合成结构自然、语义合理的篡改文本图像,使传统基于视觉痕迹或像素异常的检测与定位方法难以识别。此外,对抗样本攻击(如FGSM、BIM)可通过微小扰动误导模型做出错误判断,暴露出现有检测器在安全性方面的脆弱性。

(3)跨语种与复杂文本场景的适应性不足:多语言文本图像(如中英文混排、手写字符、少数民族语言)在字符编码、字体结构、OCR识别准确性方面存在显著差异。现有多数方法仍集中在英文数据或统一字体场景,缺乏对语言多样性和跨语种迁移学习的系统研究,限制了其在全球化文档中的实际可用性。

(4)模型可解释性与可信性缺失:当前主流方法大多依赖深层神经网络自动提取篡改特征,虽取得良好性能,但推理过程缺乏可视化和可解释输出,难以满足司法、金融等高可信场景的可审计需求。用户与

监管方难以理解模型“为何判断为篡改”,导致实际部署中存在信任障碍。

4.2 实践应用的障碍

(1)高计算资源依赖,难以移动部署:多模态融合模型(如双流 Transformer 架构)虽在准确性上表现出色,但其模型参数庞大、推理成本高,在资源受限的移动设备或边缘设备中部署存在困难。虽然已有轻量化模型探索(如 MobileNet、剪枝、蒸馏等),但其检测精度仍落后于主干模型,尚未形成完整的性能-资源折中解决方案。

(2)标准化与评估指标不统一:当前公开数据集种类繁多,注释方式(像素级、文字框级、语义级)不尽相同,评价指标(F1-score、IoU、AUC 等)标准不统一,缺乏广泛认可的跨任务基准,导致各方法间横向比较困难。部分方法在自建数据集上取得的优秀结果,难以推广至他人系统或场景。

(3)实际部署案例稀缺:尽管已有技术方案日趋成熟,但缺乏可公开验证的真实部署案例与开放测试平台。如票据审核系统、司法图像证据审查系统的应用路径、运维策略和反馈机制仍鲜有披露,限制了研究与产业的交互。

4.3 法律与监管驱动

随着篡改技术的普及与泛滥,法规和标准体系的建设已成为推动检测技术规范化的关键力量。例如:

•中国于2024年发布《文本图像篡改检测系统技术要求》,明确了检测精度、响应时间、数据安全与接口规范等技术标准;

•欧盟提出《人工智能法案》,对 AI 篡改内容的标注提出强制性要求;

•美国部分州(如加州、纽约)已立法禁止在政治广告中使用未标注的 AI 生成图像。

这些规范要求检测系统具有高准确率、可追溯性和跨平台兼容性,同时在隐私保护、数据安全、误报管理方面提出了更高标准,也对模型的可解释性和部署透明度提出挑战。

4.4 未来研究方向

结合上述问题与趋势,未来文本图像篡改检测与定位研究可从以下几个方面进一步深化。

(1)多模态融合与大型视觉语言模型结合:在现有 RGB、频域、残差等模态的基础上,进一步引入 CLIP、BLIP、GPT-4V 等大型视觉语言模型,对图像与文本进行统一表征和跨模态对齐。一方面,可利用其开放域知识提升对复杂语义篡改和跨场景文本矛盾的识别能力;另一方面,也需要系统评估和抑制大模型在篡改检测与定位中的安全风险,例如对对抗样本、分布外输入和恶意提示(prompt)的敏感性,并探索基于注意力可视化、伪造注意力图和自然语言解释等方式,提高其判别结果的可解释性和可审计性。

(2)面向对抗场景的鲁棒模型设计:在现有对抗训练、取证正则化等工作基础上,构建更加系统的“攻击—防御”框架,显式建模攻击者能力和威胁模型,设计兼顾干净样本性能与对抗鲁棒性的训练机制。未来可进一步结合多视角一致性约束、特征空间平滑正则以及随机化推理等策略,提高检测器在自然降质(压缩、模糊、重采样等)和恶意攻击(白盒、黑盒对抗样本)下的稳定性与泛化能力。

(3)跨语种、多场景迁移学习方法:针对真实应用中存在的多语言文档、手写体以及复杂场景文本等情况,有必要探索更加有效的跨语种表示学习和域自适应方法。可以利用共享子空间学习、领域对抗训练以及基于提示的参数高效微调等技术,使模型能够在不同语言、版式结构和业务场景之间实现快速迁移,缓解对大规模标注篡改数据的依赖,提升模型在全球化文档场景中的适用性。

(4)数据稀缺条件下的无监督与小样本检测:考虑到真实业务中精细标注篡改掩码的成本极高,亟需发展无监督、半监督和小样本检测策略。例如,可将文本图像篡改检测建模为异常检测问题,通过自监督学习或生成模型刻画“真实文本图像”的本征分布,从偏离该分布的区域中发现潜在篡改;也可以结合元学习、度量学习和伪样本合成,使模型在仅有极少数篡改样本的条件下仍能快速适应新场景。这类方法有望在标注极其有限甚至无标注的情况下,显著提升篡改检测的实用性。

(5)可解释性增强与人机协作检测:在司法鉴定、金融风控等高风险场景中,仅给出“篡改/未篡改”的二值判断往往难以满足可审计需求。未来可以进一步将检测结果与自然语言解释、可视化证据区域以及置信度或证据强度评分相结合,构建“模型给出候选结果—人类专家复核确认”的人机协作工作流,在保证检测准确性的同时提升决策透明度和用户信任度。

(6)低资源设备上的实时检测方案:针对移动端和边缘侧部署需求,可在保持检测性能的前提下,引入网络结构搜索、知识蒸馏、模型剪枝和量化等技术,构建轻量化、高效率的文本篡改检测模型。同时,需要在设计之初就考虑鲁棒性与安全性,避免因过度压缩模型而导致在对抗攻击或强压缩、模糊等自然降质条件下性能大幅下降。

(7)构建统一评测标准与公共平台:当前不同工作在数据集、标注粒度和评价指标上仍存在较大差异,缺乏统一的评测基准。未来应推动建立覆盖多语种、多场景、多篡改类型的公开评测平台,明确像素

级/区域级/文本级等多粒度指标以及鲁棒性指标,支持对篡改检测模型的公平、可复现比较,并促进科研界与产业界在数据、工具和典型案例层面的协同共享。

参考文献

[1] Roy P, Bag S. Detection of handwritten document forgery by analyzing writers' handwritings[C]//Pattern Recognition and Machine Intelligence. Cham: Springer, 2019: 596-605.

[2] Verdoliva L. Media forensics and DeepFakes: An overview[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(5): 910-932.

[3] 王裕鑫, 张博强, 谢洪涛, 等. 基于空域与频域关系建模的篡改文本图像检测[J]. 网络与信息安全学报, 2022, 8(3): 29-40.
Wang Yuxin, Zhang Boqiang, Xie Hongtao, et al. Tampered text detection via RGB and frequency relationship modeling[J]. Chinese Journal of Network and Information Security, 2022, 8(3): 29-40. (in Chinese)

[4] Zhao Lin, Chen Changsheng, Huang Jiwu. Deep learning-based forgery attack on document images[J]. IEEE Transactions on Image Processing, 2021, 30: 7964-7979.

[5] Luo Dongliang, Liu Yuliang, Yang Rui, et al. Toward real text manipulation detection: New dataset and new solution[J]. Pattern Recognition, 2025, 157: 110828.

[6] Lampert C H, Mei Lin, Breuel T M. Printing technique classification for document counterfeit detection[C]//2006 International Conference on Computational Intelligence and Security. Piscataway: IEEE, 2006: 639-644.

[7] 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. Piscataway: IEEE, 2018: 1053-1061.

[8] Bertrand R, Terrades O R, Gomez-Krämer P, et al. A conditional random field model for font forgery detection[C]//2015 13th International Conference on Document Analysis and Recognition. Piscataway: IEEE, 2015: 576-580.

[9] Van Beusekom J, Shafait F, Breuel T M. Text-line examination for document forgery detection[J]. International Journal on Document Analysis and Recognition (IJDAR), 2013, 16(2): 189-207.

[10] Shang Shize, Kong Xiangwei, You Xingang. Document forgery detection using distortion mutation of geometric parameters in characters[J]. Journal of Electronic Imaging, 2015, 24(2): 023008.

[11] Ryu S J, Lee H Y, Cho I W, et al. Document forgery detection with SVM classifier and image quality measures[M]// Advances in Multimedia Information Processing - PCM 2008. Berlin, HeidelbergSpringer, 2008: 486-495.

[12] Lin Zhouchen, He Junfeng, Tang Xiaoou, et al. Fast, automatic and fine-grained tampered JPEG image detection via DCT coefficient analysis[J]. Pattern Recognition, 2009, 42(11): 2492-2501.

[13] Cruz F, Sidère N, Coustaty M, et al. Local binary patterns for document forgery detection[C]//2017 14th IAPR International Conference on Document Analysis and Recognition. Piscataway: IEEE, 2017: 1223-1228.

[14] Liang Weipeng, Dong Li, Wang Rangding, et al. Robust document image forgery localization against image blending[C]//2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications. Piscataway: IEEE, 2022: 810-817.

[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. Piscataway: IEEE, 2019: 30-39.

[16] 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. Piscataway: IEEE, 2021: 14165-14173.

[17] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation[M]//Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. ChamSpringer International Publishing, 2015: 234-241.

[18] 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. Piscataway: IEEE, 2020: 4675-4684.

[19] 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.

[20] Zhang Yulan, Zhu Guopu, Wu Ligang, et al. Multi-task SE-network for image splicing localization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(7): 4828-4840.

[21] Roy P, Bhattacharya S, Ghosh S, et al. STEFANN: Scene text editor using font adaptive neural network[C]//2020

IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 13225-13234.

[22] Pérez P, Gangnet M, Blake A. Poisson image editing[C]// ACM SIGGRAPH 2003 Papers. New York: ACM, 2003: 313-318.

[23] Wu Liang, Zhang Chengquan, Liu Jiaming, et al. Editing text in the wild[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM, 2019: 1500-1508.

[24] Yang Qiangpeng, Huang Jun, Lin Wei. SwapText: Image based texts transfer in scenes[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 14688-14697.

[25] Chai Shang, Zhuang Liansheng, Yan Fengying. LayoutDM: Transformer-based diffusion model for layout generation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 18349-18358.

[26] Diaz M, Mendoza-García A, Ferrer M A, et al. A survey of handwriting synthesis from 2019 to 2024: A comprehensive review[J]. Pattern Recognition, 2025, 162: 111357.

[27] Artaud C, Sidère N, Doucet A, et al. Find it! Fraud detection contest report[C]//2018 24th International Conference on Pattern Recognition. Piscataway: IEEE, 2018: 13-18.

[28] Alibaba Cloud Comput. Softw. Co. Security AI Challenger Program[EB/OL]. (2020). https://tianchi. aliyun. com/competition/entrance/531812/introduction.

[29] Xu Wenbo, Luo Junwei, Zhu Chuntao, et al. Document images forgery localization using a two-stream network[J]. International Journal of Intelligent Systems, 2022, 37(8): 5272-5289.

[30] Wang Yuxin, Xie Hongtao, Xing Mengting, et al. Detecting tampered scene text in the wild[M]//Computer Vision - ECCV 2022. ChamSpringer Nature Switzerland, 2022: 215-232.

[31] Alibaba Cloud Comput. Softw. Co. Real-World Image Forgery Localization Challenge[EB/OL]. (2022). https://tianchi.aliyun.com/competition/entrance/531945/introduction.

[32] Alibaba Cloud Comput. Softw. Co. Detecting Tampered Text in Images Tianchi Competition[EB/OL]. (2023). https://tianchi. aliyun. com/competition/entrance/532052/introduction.

[33] Qu Chenfan, Liu Chongyu, Liu Yuliang, et al. Towards robust tampered text detection in document image: New dataset and new solution[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 5937-5946.

[34] Tornés B M, Taburet T, Boros E, et al. Receipt dataset for document forgery detection[M]//Document Analysis and Recognition - ICDAR 2023. ChamSpringer Nature Switzerland, 2023: 454-469.

[35] Dong Li, Liang Weipeng, Wang Rangding. Robust text image tampering localization via forgery traces enhancement and multiscale attention[J]. IEEE Transactions on Consumer Electronics, 2024, 70(1): 3495-3507.

[36] 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. (in Chinese)

[37] Zhang Lingzhi, Wen T, Shi Jianbo. Deep image blending[C]//2020 IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2020: 231-240.

[38] Sun Yu, Ni Rongrong, Zhao Yao. MFAN: Multi-level features attention network for fake certificate image detection[J]. Entropy, 2022, 24(1): 118.

[39] Ferrara P, Bianchi T, De Rosa A, et al. Image forgery localization via fine-grained analysis of CFA artifacts[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(5): 1566-1577.

[40] 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. Piscataway: IEEE, 2019: 9535-9544.

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

[42] Liao Xin, Chen Siliang, Chen Jiaxin, et al. CTP-net: Character texture perception network for document image forgery localization[PP/OL]. V2. arXiv (2023-08-15) [2025-09-23]. https://doi.org/10.48550/arXiv.2308.02158.

[43] 唐昊, 李泽超, 蒋鑫, 等. 基于视觉Transformer的双视图融合细粒度图像识别[J]. 软件学报, 2026, 37(5): 2286-2308.
Tang Hao, Li Zechao, Jiang Xin, et al. Dual-view fusion for fine-grained image recognition with vision transformer[J].

Journal of Software, 2026, 37(5): 2286-2308. (in Chinese)

[44] Liu Ze, Hu Han, Lin Yutong, et al. Swin transformer V2: Scaling up capacity and resolution[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 11999-12009.

[45] Dong Chengbo, Chen Xinru, Hu Ruohan, 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.

[46] Chen Zhongxi, Chen Shen, Yao Taiping, et al. Enhancing tampered text detection through frequency feature fusion and Decomposition[C]//Computer Vision - ECCV 2024. Cham: Springer, 2025: 200-217.

[47] Yu Zeqin, Li Bin, Lin Yuzhen, et al. Learning to locate the text forgery in smartphone screenshots[C]//ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2023: 1-5.

[48] Du Yuning, Li Chenxia, Guo Ruoyu, et al. PP-OCR: A practical ultra lightweight OCR system[PP/OL]. V3.arXiv (2020-10-15) [2025-09-23]. https://doi.org/10.48550/arXiv.2009.09941.

[49] Ren Ruyong, Hao Qixian, Gu Feng, et al. EMF-Net: An edge-guided multi-feature fusion network for text manipulation detection[J]. Expert Systems with Applications, 2024, 249: 123548.

[50] Qu Chenfan, Zhong Yiwu, Guo Fengjun, et al. Revisiting tampered scene text detection in the era of generative AI[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(1): 694-702.

[51] Li Zhenjiang, Sun Jingzhe, Wang Shu. Semantic-based conflict detection: Tampering detection research in bilingual scene images containing textual content[J]. Symmetry, 2025, 17(4): 536.

[52] Tornés B M, Boros E, Doucet A, et al. Detecting forged receipts with domain-specific ontology-based entities & relations[M]//Document Analysis and Recognition - ICDAR 2023. ChamSpringer Nature Switzerland, 2023: 184-199.

[53] Joren H, Gupta O, Raviv D. Learning document graphs with Attention for Image manipulation detection[C]//Pattern Recognition and Artificial Intelligence. Cham: Springer, 2022: 263-274.

[54] Guo Xiao, Song Xiufeng, Zhang Yue, et al. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector[PP/OL]. V1.arXiv (2025-03-26)[2025-09-23]. https://doi.org/10.48550/arXiv.2503.20188.

[55] Zhang Yue, Colman B, Guo Xiao, et al. Common sense reasoning for Deepfake detection[C]//Computer Vision - ECCV 2024. Cham: Springer, 2025: 399-415.

[56] Shao Huiru, Qian Zhuang, Huang Kaizhu, et al. Delving into adversarial robustness on document tampering localization[C]//Computer Vision - ECCV 2024. Cham: Springer, 2025: 290-306.

[57] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[PP/OL]. V4. arXiv (2019-09-04) [2025-09-23]. https://doi. org/10.48550/arXiv.1706.06083.

[58] Gu Jindong, Zhao Hengshuang, Tresp V, et al. SegPGD: An Effective and Efficient Adversarial Attack for Evaluating and Boosting Segmentation Robustness[C]//Computer Vision - ECCV 2022. Cham: Springer, 2022: 308-325.

[59] Zhang Hongyang, Yu Yaodong, Jiao Jiantao, et al. Theoretically principled trade-off between robustness and accuracy[C]//Proceedings of the International Conference on Machine Learning. 2019, 97: 7472-7482.

[60] Shao Huiru, Huang Kaizhu, Wang Wei, et al. Towards better robustness against natural corruptions in document tampering localization[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(1): 703-710.

[61] Zhuo Long, Tan Shunquan, Li Bin, et al. Self-adversarial training incorporating forgery attention for image forgery localization[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 819-834.

[62] Zhao Dan, Tian Xuedong. A multiscale fusion lightweight image-splicing tamper-detection model[J]. Electronics, 2022, 11(16): 2621.

[63] CASIA[EB/OL]. (2021-12-13). http://forensics.idealtest.org.

[64] 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. Piscataway: IEEE, 2006: 549-552.

[65] Jabbarlı G, Kurt M. LightFFDNets: Lightweight convolutional neural networks for rapid facial forgery detection[PP/OL]. V1.arXiv (2024-11-18)[2025-09-23]. https://doi.org/10.48550/arXiv.2411.11826.

[66] Qu Chenfan, Liu Jian, Chen Haoxing, et al. Explainable tampered text detection via multimodal large models[PP/OL]. V3.arXiv (2023-01-15)[2025-09-20]. https://arxiv.org/abs/2412.14816.

[67] Xiao Tete, Liu Yingcheng, Zhou Bolei, et al. Unified per-

ceptual parsing for scene understanding[M]//Computer Vision - ECCV 2018. ChamSpringer International Publishing, 2018: 432-448.

[68] Chen L C, Zhu Yukun, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[M]//Computer Vision - ECCV 2018. ChamSpringer International Publishing, 2018: 833-851.

[69] 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.

[70] Xie Enze, Wang Wenhai, Yu Zhiding, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers[J]. Advances in Neural Information Processing Systems, 2021, 34: 12077-12090.

[71] Cheng Bowen, Schwing A G, Kirillov A. Per-pixel classification is not all you need for semantic segmentation[C]// Proceedings of the Advances in Neural Information Processing Systems. 2021, 34: 17864-17875.

[72] Cheng Bowen, Misra I, Schwing A G, et al. Masked-attention mask transformer for universal image segmentation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 1280-1289.

[73] 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.

[74] 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.

[75] 李豪, 郝文宁, 邹世辰, 等. 基于Diffusion-Mamba和尺度不变损失的渐进式图像生成方法[J]. 电子学报, 2025, 53(9): 3384-3396.
Li Hao, Hao Wenning, Zou Shichen, et al. Progressive image synthesis method based on diffusion-mamba and scale-invariant loss[J]. Acta Electronica Sinica, 2025, 53(9): 3384-3396. (in Chinese)

作者简介

句福娇 女,1987年出生。现为北京工业大学计算机学院副教授、博士生导师。主要研究领域为深度学习,计算机视觉。
E-mail: jfj2017@bjut.edu.cn

齐光磊 男,1979年出生。现为北京邮电大学世纪学院计算机科学与技术系副教授。主要研究领域为深度学习、计算机视觉。
E-mail: qiguanglei@ccbupt.cn

张 浩 男,2000年出生。现为北京工业大学计算机学院硕士研究生。主要研究领域为深度学习、计算机视觉。
E-mail: zh2024@emails.bjut.edu.cn

王宏远 女,1988年出生。现为北京工业大学计算机学院讲师。主要研究领域为信息安全、大数据安全。
E-mail: wanghongyuan@bjut.edu.cn