基于多特征协同置信感知的深度伪造检测定位

郭祯[1]，邱润尧[1]，徐嘉[1]，刘志全[2]，马建峰[3]
（1.海南大学网络空间安全学院（密码学院），海南 海口 570228；2.暨南大学网络空间安全学院，广东 广州 510632；3.西安电子科技大学网络与信息安全学院，陕西 西安 710071）

摘要：随着人工智能生成技术的发展，伪造图像滥用引发的虚假信息传播问题日益突出。针对现有方法在复杂篡改场景下泛化能力不足、篡改区域边界定位不够精确的问题，本文提出一种基于多特征协同置信感知的深度伪造图像检测与定位方法。该方法通过构建语义—取证双流交互框架，协同利用图像语义信息与伪造痕迹信息，以提升对篡改区域的检测与定位能力。同时，引入小波域注意力强化高频篡改痕迹表征，并结合像素级门控融合与级联解码逐步细化篡改区域边界。训练阶段采用两阶段优化策略，通过任务适配器引入图像级检测与置信度估计，以缓解多任务优化干扰。实验结果表明，相较于现有代表性方法，该方法的平均F1值和平均AUC值分别达到69.5%和90.7%，在复杂篡改场景下表现出较好的检测性能与泛化能力。

关键词：深度伪造检测与定位；多特征协同；小波域注意力；级联解码
中图分类号：TP309
文献标志码：A
doi：10.11959/j.issn.1000

Multi-Feature Collaborative Confidence-Aware Method for Deepfake Detection and Localization

Guo Zhen[1], Qiu Runyao[1], Xu Jia[1], Liu Zhiquan[2], Ma Jianfeng[3]
1. School of Cyberspace Security (School of Cryptography), Hainan University, Haikou 570228, China
2. College of Cyber Security, Jinan University, Guangzhou 510632, China
3. School of Cyber Engineering, Xidian University, Xi'an 710071, China

Abstract: With the development of artificial intelligence-based image generation technologies, the misuse of manipulated images has increasingly fueled the spread of misleading information. To address the limited generalization and inaccurate boundary localization of existing methods in complex manipulation scenarios, a deepfake image detection and localization method based on multi-feature collaborative confidence awareness was proposed. The proposed method adopts a semantic-forensic dual-branch interactive framework to jointly exploit image content information and forgery trace information, thereby improving the detection and localization of tampered regions. A wavelet-domain attention module is further introduced to enhance the representation of high-frequency tampering traces, while pixel-level gated fusion and cascaded decoding are used to progressively refine manipulated boundaries. During training, a two-stage optimization strategy is adopted, where a task adapter introduces image-level detection and confidence estimation to alleviate multi-task optimization interference. Experimental results show that, compared with representative competing methods, the proposed method attains an average F1 of 69.5% and an average AUC of 90.7%, respectively. It achieves strong detection performance and generalization ability in complex manipulation scenarios.

Key words: deepfake detection and localization, multi-feature collaboration, wavelet attention, cascaded decoding

---

收稿日期：XXXX–XX–XX；修回日期：XXXX–XX–XX
通信作者：徐嘉，jiaxu@hainanu.edu.cn
基金项目：国家自然科学基金资助项目（No.62562027）
Foundation Items：The National Natural Science Foundation of China (No.62562027)

0 引言

深度学习与计算机视觉技术的持续发展推动了人工智能生成内容（Artificial Intelligence Generated Content, AIGC）的快速演进，形成了高保真、低门槛的视觉内容生成新范式。随着扩散模型与生成式编辑技术在移动终端及在线平台中的广泛部署，图像合成与内容篡改呈现出低成本、高逼真度和易传播等特征[1]，其在虚假新闻传播、舆情干预及身份冒用等场景中的滥用对社会信任体系和网络空间安全构成了严峻挑战。面对日益复杂的安全威胁，仅依赖图像级的真伪判定难以满足精准审查与责任认定的实际需求[2]。因此，在深度伪造图像检测的基础上进一步实现篡改区域的精确定位，通过生成可解释的篡改区域掩码以提供空间取证依据，已成为当前数字内容安全治理的重要研究方向。

目前，深度伪造图像检测与定位已涌现出大量基于深度学习的取证方法，并在特定数据集和应用场景中取得了一定进展。然而，在复杂退化条件与开放环境下，其稳定性、泛化性与定位精度仍然受限。Kong 等人[3]基于像素不一致性建模篡改痕迹，但在未见数据和受扰动图像上的泛化性与鲁棒性仍显不足，说明低层取证线索在压缩失真、噪声扰动等复杂退化条件下难以被稳定保持和有效表征。Su 等人[4]从非语义中心视角研究图像篡改定位，通过挖掘与图像内容弱相关的取证线索提升定位性能，说明仅依赖语义表征难以充分捕获篡改痕迹。语义表征与取证残差在尺度分布与统计特性上的显著差异，进一步限制了二者的协同建模与局部边界细节刻画。Guillaro 等人[5]通过联合建模噪声指纹、异常线索与置信信息提升篡改定位可靠性，但检测、定位与置信度的联合优化仍易受到梯度竞争与负迁移的影响，从而限制多任务协同性能。因此，如何在复杂退化环境下强化微弱取证线索表征，促进语义信息与取证残差的有效协同，并提升检测—定位联合建模的稳定性与可信性，仍是深度伪造图像取证研究亟待解决的关键问题。

针对上述问题，本文提出一种多特征协同置信感知的深度伪造图像检测与定位方法。该方法构建语义—取证双流交互框架，联合建模图像高层语义上下文和低层取证残差信息。区别于依赖全局频谱或固定频段建模的频域注意力方法，本文在语义分支中引入小波域注意力机制，在具有空间对应关系的多尺度高频子带中自适应强化异常线索，从而增强微弱频域伪影表征并保留局部结构信息。针对传统双流方法采用拼接或加和等静态融合方式易引发异构特征冲突的问题，本文设计像素级自适应门控机制，动态调节语义特征与取证特征的融合权重，并结合级联边缘细化策略提升复杂边界和微小篡改区域的定位精度。同时，结合基于任务适配器的图像级检测与置信度感知机制，增强检测、定位与置信度估计协同建模的稳定性和预测可靠性。本文的主要贡献如下：

（1）提出语义—取证双流特征协同表征框架。通过语义分支与取证分支的交互建模，实现高层语义上下文与低层取证残差信息的互补表征。引入小波域注意力机制自适应增强多尺度高频异常线索，提高模型对微弱伪造痕迹的感知能力。

（2）提出像素级自适应门控融合与级联边缘细化机制。通过像素级门控权重动态调节语义特征与取证残差的融合比例，实现异构特征的区域自适应融合。结合级联边缘细化策略逐级强化边界约束，提高复杂边界和微小篡改区域的定位精度。

（3）提出多任务软解耦与置信度感知机制。通过任务适配器对共享特征进行任务相关调节，缓解检测与定位联合优化中的梯度竞争。引入置信度感知与校准机制，以增强模型在压缩失真和噪声扰动等退化场景下的鲁棒性与泛化能力。

1 相关工作

1.1 传统图像篡改检测与定位方法

早期图像篡改取证主要基于自然图像的物理属性和统计规律，通过分析像素层面的异常特征检测潜在篡改痕迹。Li 等人[6]基于分层特征点匹配实现了复制—移动伪造检测。Korus 等人[7]提出面向光响应非均匀性篡改定位的多尺度分析方法，提高微小伪造区域的检测精度。Matern 等人[8]则通过梯度光照描述增强了光照一致性检测的鲁棒性。此类方法通常围绕特定伪造痕迹构造人工特征，具有较强可解释性，并在传统拼接与复制—移动任务中取得了较好效果。然而，该类方法通常依赖特定成像假设或预设篡改痕迹，其特征表达能力局限于显式可建模的异常模式。当图像经历复杂非线性编辑、压缩失真或跨场景分布变化时，底层统计特征易被破坏或掩盖，导致其稳定性与泛化能力显著下降，难

以适应生成式伪造场景中更为隐蔽的细粒度篡改。

1.2　基于深度空间表征的图像取证方法

随着深度学习技术的发展，卷积神经网络和Transformer等模型逐渐成为图像篡改检测与定位研究的主流方法。通过端到端学习图像中的潜在伪影特征，深度模型能够自动提取多层次判别信息，从而提升复杂场景下的检测与定位能力。早期研究主要基于卷积结构对局部异常线索进行建模，Bayar等人[9]通过约束卷积抑制语义信息以突出底层取证线索，实现对多种编辑操作的有效检测。Bappy等人[10]结合重采样特征与长短期记忆网络，实现了像素级篡改区域定位。这类方法在局部纹理异常与残差特征建模方面具有较强能力。然而，受限于卷积操作的局部感受野，模型难以有效捕获长距离依赖与全局语义一致性。当伪造区域与真实背景在局部纹理上高度相似时，模型易产生误判或定位漂移。

为增强全局建模能力，近年来研究逐渐引入Transformer结构与大规模视觉预训练模型，通过自注意力机制建立跨区域依赖关系，从而增强对全局语义一致性的建模能力。Hao等人[11]提出TransForensics网络，通过密集自注意力建模多尺度图像块间的依赖关系。Wang等人[12]提出ObjectFormer网络，将对象级表示与全局语义建模引入篡改检测与定位任务。总体而言，基于深度空间表征的方法虽在全局关系建模和语义不一致性刻画方面具有优势，但其特征表达更偏向高层语义信息，对微弱取证线索（如高频残差与细粒度伪影）的感知能力有限，在复杂退化和跨域条件下仍易出现检测偏差或边界模糊。

1.3　基于深度频域建模的图像取证方法

考虑到基于空间表征的方法对微弱伪影的感知能力有限，相关研究逐渐将频域信息引入深度图像取证框架。Qian等人[13]提出F3-Net，通过频率分解与局部频率统计对伪造模式进行建模，在压缩场景下提升检测性能。Liu等人[14]利用频域相位信息增强模型在跨数据集与压缩条件下的鲁棒性。Luo等人[15]引入高频噪声线索，以缓解模型对特定颜色纹理的过拟合。近年来，Tan等人[16]与Li等人[17]分别从联合建模与数据增强角度探索频域信息的利用方式，相关研究逐渐由单一频域特征利用转向频域与空间表征的协同建模。然而，现有方法多基于整体频谱进行特征提取，通常将频域信息作为辅助分支或附加特征使用，缺乏对不同频率成分重要性的精细建模。此外，频域特征与高层语义信息之间缺乏有效协同机制，在复杂退化条件下仍难以稳定刻画细粒度篡改伪影。

1.4　联合建模与多任务协同优化方法

在上述空间表征与频域建模方法的基础上，图像篡改取证研究逐渐由单一检测或定位任务转向图像级判别、像素级定位与结果解释的联合建模。Huang等人[18]提出SIDA，在统一框架下实现图像级检测、区域定位与文本解释，通过多任务协同提升模型判别能力与结果可解释性。Yu等人[19]提出PCGrad，通过梯度投影缓解不同任务间的梯度冲突。Shi等人[20]提出渐进式参数共享策略，通过控制任务间共享程度以提升多任务训练的稳定性。在图像篡改取证场景中，图像级检测更依赖全局语义信息，而像素级定位更依赖局部取证残差，两者在特征需求与优化目标上存在显著差异。现有方法虽在一定程度上利用了任务间的互补信息，但在语义特征与取证特征的分工建模以及多任务干扰抑制方面仍缺乏有效的任务协调机制，导致模型在判别能力、定位精度与结果可靠性之间难以取得稳定平衡。

针对上述问题，本文提出一种基于多特征协同置信感知的深度伪造图像检测与定位方法，通过构建语义—取证双流交互机制、引入频带级小波增强以及采用任务适配与解耦优化策略，提升复杂退化场景下的检测鲁棒性与篡改区域定位精度。

2　系统模型

2.1　总体架构

为了实现对深度伪造图像的精确检测与篡改区域定位，本文提出一种基于多特征协同与置信度感知的深度伪造图像检测与定位模型，其整体结构如图1所示。模型整体架构由四个阶段组成：多分支特征提取、多模态特征融合、特征重建和多任务解码。在多分支特征提取阶段，构建语义分支与取证分支的双流编码结构，分别提取图像的多尺度语义特征与高频取证特征。多模态特征融合阶段通过特征对齐模块（feature alignment, FA）与像素级自适应门控融合模块（gated fusion, GF）实现语义信息与取证残差特征的动态融合，从而强化微弱篡改线

索的判别表征，提高定位掩码的空间分辨能力。特征重建阶段采用级联渐进式解码结构逐步恢复特征的空间分辨率，以增强篡改区域的结构表达。多任务解码阶段引入任务适配器与置信度感知机制以缓解联合训练中的梯度竞争，

在保持像素级定位性能的同时实现图像级伪造判别与预测不确定性建模，从而提升模型在压缩失真与噪声扰动等退化场景下的鲁棒性与泛化能力。

2.2　多分支特征提取模块

深度伪造图像中的篡改线索通常同时体现在高层语义一致性与底层统计异常两个层面。为同时刻画全局语义结构与局部微弱篡改痕迹，本文构建由语义分支与取证分支组成的多分支特征提取模块，通过协同建模语义上下文信息与高频取证残差，实现对伪造痕迹的多粒度表征。

2.2.1　多尺度语义编码

语义分支采用预训练 Transformer 模型 DINOv2 ViT-L/14 作为特征提取骨干。相比依赖图文对齐监督的预训练模型，DINOv2 基于大规模自监督学习，能够在无需显式语义标签的情况下学习稳定的结构表示，并同时建模图像级与块级特征。对于图像篡改定位任务，仅依赖局部异常检测往往不足，还需结合场景语义与区域关系，对可疑区域与真实内容之间的一致性进行建模。因此，DINOv2 提供的层次化图像块表示更适用于捕获跨区域依赖与结构关系，在复杂场景下具有更强的语义建模能力与跨域泛化潜力。其冻结特征具备良好的迁移性，可降低高层语义表示从头学习的难度，并为后续多尺度融合提供层次清晰的中间表示。

设输入图像为 $I \in \mathbb{R}^{H \times W \times 3}$，从网络第 5，11，17，23 层提取中间特征构成层级集合 $L \in \{l_1,l_2,l_3,l_4\}$，对应特征表示为 $\{S_l\}_{l \in L}$。为统一不同层级特征的空间尺度，引入特征金字塔网络（feature pyramid network, FPN）进行多尺度特征融合，其结构如图 2 所示。通过侧向映射 $\phi_l(\cdot)$ 对各层特征进行通道对齐，并采用自顶向下的上采样策略传播高层语义信息，同时利用卷积映射 $\varphi_l(\cdot)$ 消除混叠效应。最终将各层级特征对齐并拼接，通过 1 × 1 聚合映射函数 $\eta(\cdot)$ 得到包含全局语义上下文与细粒度结构信息的统一语义特征图 F_s

$$F_s = \eta\left(\mathrm{Concat}\left(\left[\,\mathrm{Up}\left(\phi_l(S_l)\right)\right]_{l \in L}\right)\right) \tag{1}$$

然而，基于空间表征的方法在特征聚合过程中通常更偏向语义一致性建模，对细粒度篡改伪影的敏感性仍然有限。为增强语义特征中与篡改相关的高频响应，本文在语义分支中引入小波域注意力机制，其结构如图 3 所示。与现有直接对原始图像进行频域建模的方法不同[21-22]，本文在深层语义特征空间中进行频率分解。原始图像的频率分量主要反映像素强度变化，而经过多层 Transformer 编码后的语义特征，其高频分量更多对应特征响应的空间变化模式，即区域间的不连续性、结构错位以及边

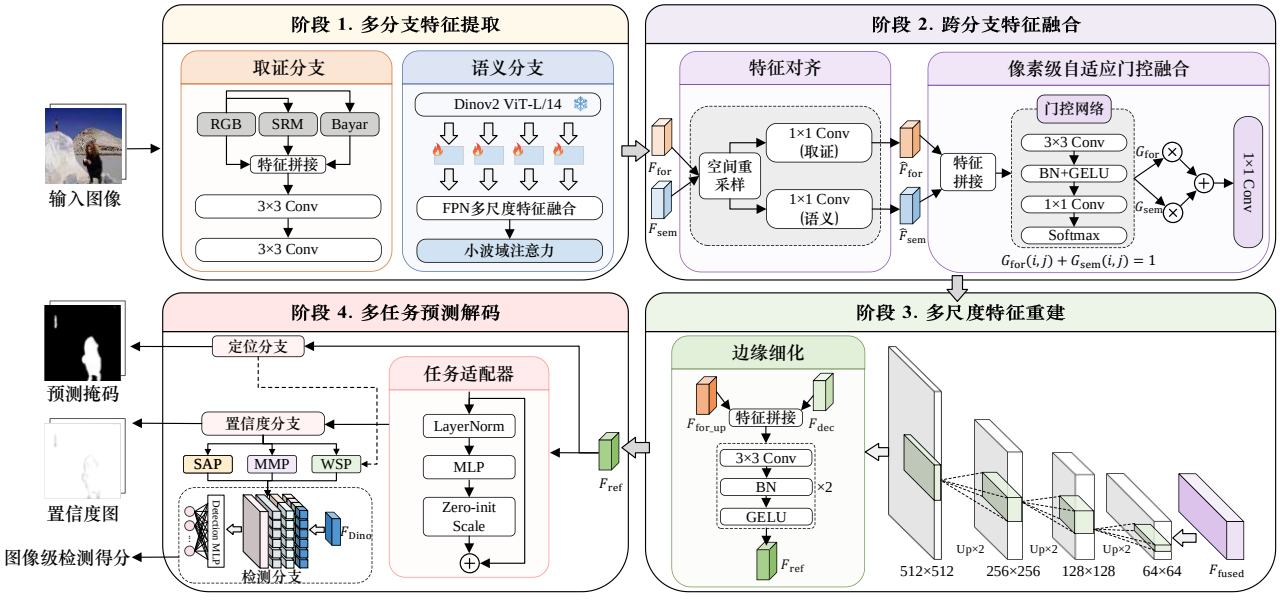


图 1　基于多特征协同置信感知的深度伪造图像检测与定位框架

界过渡。这类变化在篡改区域附近通常更为显著，因而具有更强的判别能力。

从数学机理上看，离散小波变换（discrete wavelet transform，DWT）具有多分辨率分析与时频局部化特性。其本质是将输入特征投影到一组具有尺度和平移属性的正交小波基函数上，从而在保持空间对应关系的同时，将特征分解为低频近似分量和不同方向的高频细节分量。其中，低频分量主要保留整体语义结构和上下文信息，高频分量则分别刻画水平、垂直和对角方向上的局部变化。相较于全局频谱建模或空域统一加权注意力机制，小波分解能够从方向维度分离局部结构变化，从而增强对边界过渡、区域不连续和细粒度扰动等篡改线索的表征能力。因此，本文通过对高频子带进行自适应重加权，在增强微弱伪影表征的同时保留其空间位置关系，为篡改区域定位提供更具判别性的特征表示。具体地，本文对语义特征 F_s 进行单级小波分解，得到低频分量与三个方向的高频分量为

$$\{B_{\mathrm{LL}},B_{\mathrm{LH}},B_{\mathrm{HL}},B_{\mathrm{HH}}\}=\mathrm{DWT}(F_s) \tag{2}$$

其中，B_{LL} 为低频近似分量，高频分量 $B_{\mathrm{LH}},B_{\mathrm{HL}},B_{\mathrm{HH}}$ 分别刻画水平、垂直与对角方向的局部变化。随后，通过卷积映射生成高频注意力权重 A_k，并对高频子带进行逐元素加权

$$B'_k=A_k\odot B_k,k\in\{\mathrm{LH},\mathrm{HL},\mathrm{HH}\} \tag{3}$$

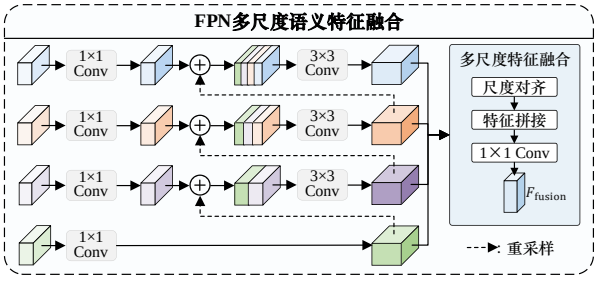


图2　基于FPN的语义特征多尺度融合模块

低频子带 B_{LL} 仅通过轻量卷积映射 $\phi_{\mathrm{lf}}(\cdot)$ 完成特征变换，以保留深层特征中的全局语义与结构上下文信息。在此基础上，借助逆离散小波变换（inverse discrete wavelet transform, IDWT）将低频分量与加权后的高频分量重构为空间域特征，并通过残差连接保留原始语义信息，最终得到增强后的特征表示为

$$F_{\mathrm{sem}}=F_s+\mathrm{IDWT}\left(\phi_{\mathrm{lf}}(B_{\mathrm{LL}}),B'_k\right) \tag{4}$$

本文将小波域注意力作用于融合后的深层语义特征，而非浅层特征。浅层特征虽包含更丰富的高频细节，但同时混入较多成像噪声、压缩扰动及无关纹理。相比之下，深层语义特征在上下文约束下，其高频响应更稳定地反映结构失配与边界异常。因此，该设计在保持语义一致性的同时强化了对细粒度结构变化的刻画能力，为篡改区域定位提供更具判别性的特征表示。

2.2.2　自适应取证编码

取证分支侧重于捕获图像中的统计异常与纹理不连续等物理层伪造线索。为充分保留底层篡改痕迹并兼顾不同篡改模式的表征需求，本文构建由原始图像、固定空间富模型（spatial rich model, SRM）以及可学习的Bayar约束残差组成的多路径取证编码结构，在网络浅层显式引入对边界变化与统计异常更为敏感的高频响应，以加强对微弱篡改痕迹的表征。

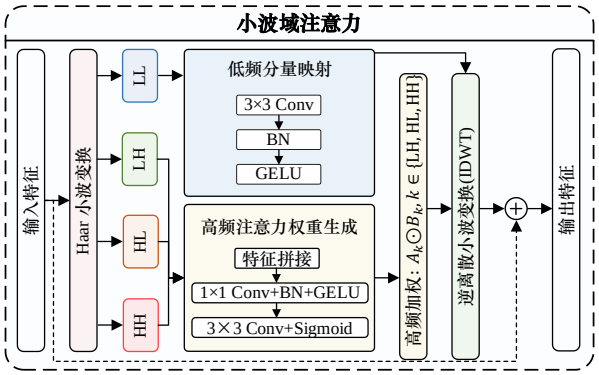


图3　小波域注意力模块

设输入图像为 I，原始RGB分支用于提供整体结构与内容信息。SRM分支采用3组经典5 × 5高通滤波器对输入图像进行分组卷积，每组包含3个滤波核，共生成9通道固定残差特征 F_{srm}。卷积步长设为1，padding设为2以保持空间分辨率不变。该分支参数保持冻结，主要用于稳定提取边缘突变、噪声不一致和局部统计异常等高频取证线索。

为提升模型对复杂篡改模式的适应能力，本文进一步引入Bayar约束卷积分支生成自适应残差特征 F_{bayar}。具体地，采用12个可学习的5 × 5约束卷积核，并对卷积核 W 施加零直流响应约束，压制内容分量的同时保留高频残差。设卷积核中心位置为 c，其约束条件为 $\sum_{i\neq c}W_i=1,W_c=-1$。该约束使卷积

操作等价于对中心像素进行邻域预测后的残差表达，其输出可表示为

$$y = I \times W = \sum_{t \neq c} w_t I_t - I_c \tag{5}$$

对于符合自然成像统计规律的区域，预测残差趋近于零，而在篡改区域，由于统计特性发生改变，预测误差显著增大。基于上述约束，Bayar分支能够在训练过程中根据不同篡改操作引起的局部统计差异自适应调整卷积权重，从而提取更具判别性的残差特征。从特征作用上看，SRM分支提供稳定且可解释的统计异常响应，Bayar分支通过可学习约束提取面向不同篡改模式的自适应残差特征。前者有助于保留通用高频线索，后者能够补充固定滤波器难以覆盖的复杂异常模式，二者在特征空间中形成互补。

最终，将原始图像特征、SRM特征与Bayar特征在通道维度进行拼接，并通过卷积融合模块映射至统一特征空间，得到高分辨率取证特征

$$F_{\text{for}} = \text{Conv}\left(\text{Concat}\left(I, F_{\text{srm}}, F_{\text{bayar}}\right)\right) \tag{6}$$

该取证编码模块在保持统计异常响应稳定性的同时，引入对复杂篡改模式的自适应刻画能力，为后续语义—取证协同提供可靠的高频残差特征。

2.3 多模态特征融合模块

语义特征F_{sem}与取证特征F_{for}在表示形式与空间分布上存在显著差异。前者侧重全局语义一致性建模，后者强调局部统计异常响应。直接采用线性融合或简单拼接难以有效刻画二者的互补关系，甚至可能导致语义信息掩盖细粒度篡改痕迹。为实现语义信息与取证线索的有效协同，本文设计像素级自适应GF，通过位置相关的动态权重分配机制实现双分支特征的交互建模。

由于两类特征在空间分辨率与通道表达上存在差异，需先经线性投影与空间重采样进行特征对齐。以语义分支特征F_{sem}的空间尺度为基准，当两路特征分辨率不一致时，采用双线性插值对F_{for}进行重采样，使其与F_{sem}具有一致的空间尺寸。随后，通过两组独立的1×1卷积对两路特征进行通道映射，得到对齐后的语义特征$\hat{F}_{\text{sem}}$和取证特征$\hat{F}_{\text{for}}$。将对齐后的特征沿通道维度拼接，门控生成函数$\psi(\cdot)$采用轻量卷积结构实现，其具体形式为

$$\psi(F) = \text{Conv}_{1\times1}\left(\text{GELU}\left(\text{BN}\left(\text{Conv}_{3\times3}(F)\right)\right)\right) \tag{7}$$

通过Softmax在通道维度生成逐位置归一化权重

$$\left[g_{\text{sem}}, g_{\text{for}}\right] = \text{Softmax}\left(\psi\left(\text{Concat}\left(\hat{F}_{\text{sem}}, \hat{F}_{\text{for}}\right)\right)\right) \tag{8}$$

其中，g_{sem}与g_{for}分别表示语义特征和取证特征在各空间位置上的权重，并满足每个空间位置上$g_{\text{sem}} + g_{\text{for}} = 1$。基于上述权重，对两路特征进行逐像素加权融合，并经输出映射函数$\Phi(\cdot)$得到最终融合特征为

$$F_{\text{fused}} = \Phi\left(g_{\text{sem}} \odot \hat{F}_{\text{sem}} + g_{\text{for}} \odot \hat{F}_{\text{for}}\right) \tag{9}$$

与固定加权或简单拼接方式不同，该机制通过空间自适应权重分配实现语义信息与取证线索的协同建模。具体而言，在结构平滑区域，语义特征更为稳定，较大的g_{sem}有助于抑制取证分支引入的噪声响应。而在篡改边界或重采样区域，统计异常更为显著，较大的g_{for}有助于保留残差信号中的有效取证线索。

多模态特征融合模块通过位置相关的动态权重调节实现双流特征交互。语义特征提供全局结构约束，取证特征提供局部异常响应，门控权重用于动态平衡二者贡献，从而在全局一致性与局部判别性之间取得平衡，使融合特征同时具备语义结构完整性与局部异常敏感性，进而提升复杂退化条件下的篡改定位性能。

2.4 特征重建模块

语义—取证融合特征通常处于低分辨率空间，直接上采样易导致篡改边界模糊及细节信息丢失。为恢复空间结构并提升边界刻画能力，本文采用三级级联解码结构进行特征重建，并在高分辨率阶段引入边缘细化模块（edge refinement, ER）以增强局部细节信息。设融合特征F_{fused}为解码器输入，第l级解码过程可表示为解码过程表示为

$$F_{\text{dec}}^{(0)} = F_{\text{fused}}, F_{\text{dec}}^{(l)} = D_l\left(F_{\text{dec}}^{l-1}\right), l \in \{1,2,3\} \tag{10}$$

其中，$D_l(\cdot)$表示第l级解码单元，其结构形式为

$$D_l(\cdot) = \text{GELU}\left(\text{BN}\left(\text{Conv}\left(\text{Deconv}(\cdot)\right)\right)\right) \tag{11}$$

其中，$\text{Deconv}(\cdot)$表示转置卷积上采样操作。每一级解码在提升空间分辨率的同时对局部特征进行重组，从而逐步恢复结构信息。

考虑到篡改区域的边界定位对高频信息尤为敏感，在最高分辨率阶段额外引入ER。该模块以解

码特征F_{dec}与取证分支输出的高频残差特征F_{for}为输入。由于两类特征在空间尺寸上可能存在差异，首先通过双线性插值对将F_{for}进行特征对齐，得到$F_{\text{for_up}}$。随后将其与F_{dec}在通道维度拼接，并通过边缘映射函数$\Psi(\cdot)$进行融合，得到最终重建特征为

$$F_{\text{ref}} = \Psi\left(\text{Concat}\left(F_{\text{dec}}, F_{\text{for\_up}}\right)\right) \tag{12}$$

其中，$\Psi(\cdot)$由卷积、归一化与非线性激活构成，用于增强边界过渡与局部结构不连续信息。特征重建模块通过三级解码逐步恢复空间结构，并结合取证残差进行边界补偿，使重建特征在保持语义一致性的同时增强对细粒度篡改边界的刻画能力，从而提升复杂退化条件下的像素级定位性能。

2.5　多任务解码模块

在篡改检测任务中，像素级定位与图像级判别通常共享同一特征表示，但两类任务在优化目标上存在显著差异。定位任务依赖局部结构细节，而检测任务更关注全局语义一致性与统计分布特征。直接共享特征易引入梯度冲突，从而影响联合优化效果。考虑到完全共享难以适应任务差异，而完全分离又会破坏通用篡改表征，本文引入基于任务适配器的特征调制机制，在共享主干特征表示的基础上对特征施加轻量任务相关偏移，使不同任务在公共表示空间中形成适度分化，从而实现软解耦建模。具体地，针对重建特征F_{ref}，任务适配器以残差形式生成任务特定表示

$$F_{\text{adapt}} = \text{LN}\left(F_{\text{ref}} + s \cdot \text{MLP}\left(F_{\text{ref}}\right)\right) \tag{13}$$

其中，$\text{LN}(\cdot)$表示层归一化操作，$\text{MLP}(\cdot)$为两层前馈网络，s为初始化为0的可学习缩放因子。该零初始化策略保证在训练初期$F_{\text{adapt}} \approx F_{\text{ref}}$，从而维持原始特征分布的稳定性，并随着训练逐步引入任务相关差异，有效缓解多任务优化的梯度冲突。

在此基础上，定位分支直接基于F_{ref}进行像素级定位。由于压缩失真、噪声扰动和重采样等复杂退化因素会削弱局部边界响应，并可能引入与篡改无关的伪异常纹理，像素级定位结果往往存在空间不确定性。对于图像级检测任务，若对所有空间位置进行无差别聚合，将难以区分不同区域的判别可靠性，低置信度区域中的噪声响应和不确定预测容易干扰全局判别表示，从而降低检测结果的稳定性。为此，本文在检测分支中引入置信度感知策略与统计建模方法，对局部预测的空间可靠性进行显式刻画，并以此约束全局特征聚合过程，从而增强了模型对篡改区域空间分布和预测不确定性的表征能力。

具体地，由F_{adapt}生成与定位掩码同尺度的像素级置信度图C。从理论上看，C可视为对不同空间位置预测可靠性的估计，用于刻画局部取证线索对图像级判别结果的可信贡献。高置信度区域通常对应结构异常更明确、预测一致性更强的位置。低置信度区域则更易受到边界模糊、压缩噪声或背景纹理的干扰，具有更高的不确定性。不同于全局平均池化和最大池化等静态聚合方式，置信度感知聚合不再假设所有空间位置具有相同的判别价值，而是依据局部预测可靠性自适应调节不同区域的贡献权重。通过该空间可靠性约束，检测分支能够增强高可信异常区域对全局判别表示的贡献，并抑制低置信度区域中噪声响应的干扰，从而建立局部结构异常与图像级检测结果之间的稳定关联，提升复杂退化场景下检测结果的稳定性与可信性。

在置信度引导下，检测分支构建由语义特征与统计特征组成的全局表示。首先，将DINOv2提取的全局语义特征映射为128维向量F_{dino}，用于刻画整体语义一致性。随后，基于置信度加权的特征响应构建多维统计表示，以描述篡改区域的空间分布特性。具体而言，通过置信度加权的空间注意力池化（spatial attention pooling, SAP）得到特征$F_{\text{sap}} \in \mathbb{R}^4$，其由注意力加权均值、全局最大值、全局最小值及全局均值组成，用于刻画响应强度及整体异常分布。通过多尺度极值统计（multi-scale max pooling, MMP）得到特征$F_{\text{mmp}} \in \mathbb{R}^5$，包括全局最大响应、局部区域最大值的均值与标准差以及跨尺度响应差异，用于描述篡改响应在不同空间尺度下的变化特性。通过对置C与掩码差异图$D = \text{clamp}(M - 0.5, -0.5, 0.5)$的加权统计（weighted statistical pooling, WSP）得到特征$F_{\text{wsp}} \in \mathbb{R}^8$，其包含加权均值、加权方差及软极值等统计量，用于刻画置信度分布范围、边界偏移及局部波动信息。上述统计特征分别从强度、尺度与分布三个角度对篡改响应进行刻画，从而弥补单一全局特征难以描述复杂空间结构的问题。最终，将语义特征与统计特征拼接形成全局判别向量，并输入分类器得到图像级预测结果。

$$y_{\text{cls}} = \text{MLP}_{\text{cls}}\left(F_{\text{dino}},F_{\text{sap}},F_{\text{mmp}},F_{\text{wsp}}\right) \tag{14}$$

其中，总特征维度为145。多任务解码模块通过任务适配器实现共享特征的软解耦，并结合置信度引导的统计建模策略，使检测分支能够同时刻画全局语义异常与局部空间分布特征。该机制在不引入额外复杂结构的前提下，有效缓解多任务优化冲突，并提升图像级篡改检测在复杂退化与跨域场景下的鲁棒性与泛化能力。

2.6　两阶段渐进式训练策略

为减弱像素级定位与图像级判别在联合训练中的相互干扰，本文采用两阶段渐进式训练策略。第一阶段聚焦篡改区域定位，学习稳定的像素级表征。在此基础上，第二阶段引入置信度校准与图像级判别优化，在保持定位能力的同时增强模型对预测可靠性和全局判别信息的表征能力。

2.6.1　定位优先的特征学习

篡改区域通常呈小目标分布且边界模糊，单一损失函数难以兼顾区域一致性与边界结构的约束需求。本文从像素分类、区域重叠和边界结构构建联合损失函数，对定位分支施加多粒度监督，即

$$\mathcal{L}_{\text{S1}} = \mathcal{L}_{\text{Focal}} + \mathcal{L}_{\text{Dice}} + \lambda_b\mathcal{L}_{\text{Bound}} + \lambda_e\mathcal{L}_{\text{Edge}} \tag{15}$$

权重系数λ_b和λ_e用于平衡区域一致性约束与边界结构约束的作用强度。考虑到篡改区域通常面积较小且边界信息更具判别性，适当增强边界相关损失有助于提升模型对细粒度结构变化的敏感性。同时，为避免边界约束过强引起区域预测不稳定，通过权重调节实现不同损失项之间的协调。上述参数通过验证集调优获得，具体分析见实验部分。

为缓解篡改区域与背景之间的类别不平衡，本文引入焦点损失。该损失利用聚焦因子减弱易分类样本的影响，加强对困难样本的学习，其表达式为

$$\mathcal{L}_{\text{Focal}} = -\frac{1}{\sum M}\sum_{i\in\Omega} M_i\alpha_t\left(1 - p_{t,i}\right)^r \log\left(p_{t,i}\right) \tag{16}$$

其中，$p_{t,i}$表示像素i的预测概率，α_t为类别平衡因子，r为聚焦参数，M表示有效像素掩码。为约束篡改区域的整体预测结果，引入Dice损失，以刻画预测掩码与真实掩码的区域重叠关系，其定义为

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum P_iY_i + \varepsilon}{\sum P_i + \sum Y_i + \varepsilon} \tag{17}$$

其中，P表示预测概率图，Y为真实掩码，ε为平滑项。篡改边界处往往伴随细粒度结构变化，进一步引入边界加权损失以增强边界监督。通过对真实掩码进行形态学膨胀与腐蚀操作构建边界区域，并据此定义权重$w_i = 1 + \eta B_i(Y)$。边界损失定义为

$$\mathcal{L}_{\text{Bound}} = -\frac{\sum_i w_i\mathcal{L}_{\text{BCE}}\left(P_i,Y_i\right)}{\sum_i w_i + \varepsilon} \tag{18}$$

其中，$\mathcal{L}_{\text{BCE}}(\cdot)$表示逐像素二元交叉熵损失，$\eta$为边界增强系数。此外，为约束预测结果在边缘结构上的一致性，引入边缘一致性损失，其表达式为

$$\mathcal{L}_{\text{Edge}} = \left\|\varepsilon(P) - \varepsilon(Y)\right\|_1 \tag{19}$$

其中，$\varepsilon(\cdot)$表示通过Sobel算子提取的边缘响应。上述四项损失分别从像素分布、区域重叠及边界结构三个层面约束定位结果，协同作用下使模型获得鲁棒的像素级定位能力。

2.6.2　置信度校准与协同决策优化

第二阶段在第一阶段训练所得参数基础上继续优化。为避免对已学习的定位能力产生干扰，冻结主干网络及定位分支参数，仅更新置信度分支与图像级检测分支，使模型在保持像素级定位精度的同时学习预测可靠性与全局判别信息。考虑到实际场景中图像常伴随多种质量退化，训练过程中对输入施加噪声叠加、高斯模糊、压缩失真及随机遮挡等增强操作，以扩大样本的退化分布范围，从而促使模型建立预测不确定性与图像退化程度之间的对应关系。该阶段的联合优化损失函数定义为

$$\mathcal{L}_{\text{S2}} = \lambda_c\mathcal{L}_{\text{Conf}} + \lambda_d\mathcal{L}_{\text{Det}} \tag{20}$$

其中，$\mathcal{L}_{\text{Conf}}$为置信度校准损失，$\mathcal{L}_{\text{Det}}$为图像级检测损失。为刻画定位结果的可靠程度，引入真实类概率（true class probability, TCP）作为置信度监督信号。设第一阶段定位分支输出的概率图为P。通过对真实掩码Y进行腐蚀得到确定伪造区域Ω_f，对Y膨胀后取反得到确定真实区域Ω_r，并定义有效监督区域为$\Omega_v = \Omega_f \cup \Omega_r$。在该区域内，TCP监督目标定义为

$$C_i^* = \begin{cases} \hat{p}_i, & i \in \Omega_f \\ 1 - \hat{p}_i, & i \in \Omega_r \end{cases} \tag{21}$$

其中，$\hat{p}_i$表示像素i的预测概率。该置信度监督并非来源于独立标注，基于第一阶段预测的TCP构造得到。由于缺乏逐像素置信度真值，本文采用该自监督近似以刻画预测概率与其可靠性之间的对应关系。与直接将预测概率作为置信度不同，该策略

通过限制监督区域，仅在高置信的确定区域内进行约束，从而减弱噪声标签对学习过程的干扰。

该方法仍依赖于第一阶段预测结果，因此可能存在误差传递问题，尤其在边界过渡区域，其置信度估计仍可能受到预测偏差的影响。尽管如此，该近似监督能够提供稳定的相对置信度信号，并与退化增强策略相结合，有助于提升模型对不确定性的刻画能力。基于上述监督，置信度分支输出逐像素的置信度图C，并通过回归预测置信度与TCP目标的差异进行优化。其损失函数定义为

$$\mathcal{L}_{\text{Conf}} = \frac{\sum_{i \in \Omega_v} \left(C_i - C_i^* \right)^2}{\left| \Omega_v \right| + \varepsilon} \tag{22}$$

为提升模型对图像整体篡改的判别能力，在定位特征基础上引入图像级检测分支，其损失定义为

$$\mathcal{L}_{\text{Det}} = -\beta y \log\left(\sigma\left(\hat{y}\right)\right) - \left(1 - y\right)\log\left(1 - \sigma\left(\hat{y}\right)\right) \tag{23}$$

其中，$\hat{y}$为检测分支输出，y为图像级真实标签，β为正样本权重，用于缓解正负样本不平衡问题。经上述联合优化，主干参数冻结，梯度仅回传至置信度与检测两个分支，在不破坏已有定位能力的前提下完成置信度估计与图像真伪判别的协同建模。

3　实验与结果分析

3.1　实验设置

3.1.1　实验数据集

训练集与测试集的具体信息如表1所示。为评估所提方法在复杂篡改场景下的有效性与泛化能力，本文构建了由多个公开数据集组成的混合训练数据集。包括Fantastic-Reality[23]、IMD2020[24]、CASIA v2[25]以及文献[26]提出的SP-COCO、CM-COCO和JPEG-RAISE数据集。上述数据集覆盖了拼接、复制—移动、目标移除及混合篡改等典型篡改类型，为模型学习多样化的篡改模式提供了充分的样本基础。

考虑到不同数据源在规模与类别分布上的显著差异，尤其是SP-COCO/CM-COCO数据量远大于其他数据集，若直接使用全部样本进行训练，模型可能偏向于学习该类篡改特征。为避免这一问题，训练阶段采用基于最小子集截断与索引映射调度的类平衡采样策略。具体地，在每个训练周期中，从各子数据集中按相同数量随机抽取样本，并通过索引映射机制实现跨轮次的均匀遍历，从而保证不同篡改类型在训练过程中的出现频率基本一致。该策略在保持数据多样性的同时，有效抑制了大规模数据集对训练过程的主导作用，从而提升了模型对不同篡改类型的均衡建模能力。最终训练集包含19,020个样本，验证集1,500个样本。

表1　实验数据集具体信息

数据集	真实图片/张	篡改图片/张	主要篡改类型
训练数据集			
Fantastic Reality	16000	16000	混合篡改
IMD2020	2010	2010	混合篡改
CASIA v2	7491	5123	复制—移动、拼接
SP COCO/CM COCO	-	~800000	复制—移动、拼接
JPEG RAISE	24462	-	真实图像
测试数据集			
CASIA v1	800	921	拼接
MISD	618	300	多重拼接
Coverage	100	100	复制—移动
Columbia	183	180	拼接
NIST16	560	564	复制—移动、修复
CocoGlide	512	512	扩散模型局部编辑

3.1.2　评价指标

为全面评估模型的性能，本文分别采用不同指标进行评价。像素级篡改定位任务采用F1分数和交并比（intersection over union, IoU）作为评价指标。F1分数综合考虑精确率与召回率的平衡关系，能够从整体上反映篡改区域的检测质量，定义为

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{24}$$

其中，TP，FP，FN和TN分别表示真阳性、假阳性、假阴性和真阴性。IoU衡量预测篡改区域P与真实篡改区域Y的重叠程度，其定义为

$$\text{IoU} = \frac{P \cap Y}{\left| P \cup Y \right|} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{25}$$

图像级篡改检测任务采用ROC曲线下面积（area under curve, AUC）和等错误率（equal error rate, EER）作为评价指标。AUC用于衡量模型在不同判别阈值下区分篡改图像与真实图像的综合能力，数值越高表示判别性能越好。EER用于衡量模型在两类错误达到平衡时的检测性能，即当假阳性率（false positive rate, FPR）与假阴性率（false

negative rate, FNR）相等时对应的错误率，定义为

$$\mathrm{EER} = \mathrm{FPR}(\tau^*) = \mathrm{FNR}(\tau^*) \tag{26}$$

其中，τ^*表示使假阳性率与假阴性率相等时的判别阈值。EER越低，说明模型对真实样本与伪造样本的区分能力越强。

3.1.3 实验环境与参数设置

本文实验在配备Intel Core i9-14900K CPU和NVIDIA GeForce RTX 5090D GPU的设备上完成。基于Python 3.10和PyTorch 2.7深度学习框架实现。在训练策略上，采用两阶段渐进式优化方案。第一阶段使用AdamW优化器进行主干特征学习，通过引入解耦权重衰减以抑制过拟合并提升特征泛化能力。第二阶段冻结主干网络参数，仅对置信度分支与检测分支进行优化，并采用Adam优化器以适应较小规模参数空间的稳定更新。由于两阶段优化目标与参数空间不同，优化器切换不会引入训练不稳定性。初始学习率设为$\eta_0 = 1 \times 10^{-4}$，并采用带下界截断的多项式衰减策略，在训练后期保留一定的梯度更新能力，以避免模型在复杂非凸损失曲面过早收敛，从而提升对细粒度篡改线索的表征能力。两阶段训练过程中涉及的输入尺寸、批次大小、训练轮数、优化器及学习率等关键超参数设置如表2所示。

表2 模型训练超参数设置

超参数	阶段一	阶段二
输入尺寸	448 × 448	448 × 448
批次大小	8	8
最大训练轮数	50	50
早停轮数	5	5
优化器	AdamW	Adam
初始学习率	1×10^{-4}	5×10^{-5}
学习率调度策略	多项式衰减（p=0.9）	多项式衰减（p=0.9）

3.2 损失函数权重实验结果分析

为分析阶段一训练中各损失项权重系数λ_b和λ_e对篡改定位性能的影响，本文在CASIA v1[25]、MISD[36]、Coverage[37]、Columbia[38]、NIST16[39]和CocoGlide[5]六个数据集上针对不同权重组合进行了参数消融实验，结果如表3所示。

从整体趋势来看，模型性能对两类权重均较为敏感，且二者存在明显的耦合关系。当λ_e取值过大时，多数组合下的平均性能均出现下降，说明过强的边界约束会强化对局部高频结构的关注，从而在一定程度上干扰对主体篡改区域的整体建模。相反，当λ_e取值过小时，边界信息约束不足，容易导致定位结果边界模糊。对于区域相关权重λ_b，过小会削弱区域一致性约束，使模型难以形成完整的篡改区域预测；而过大则可能抑制边界细节的刻画能力，导致预测结果趋于过度平滑。因此，两类权重需要在区域一致性与边界结构约束之间取得平衡。

从具体结果来看，适中的权重配置通常能够获得更优且更稳定的性能。当$\lambda_b = 0.3, \lambda_e = 0.3$时，模型的平均F1值达到68.2%，平均IoU达到60.2%，优于大多数其他配置。在此基础上进一步比较可见，当$\lambda_b = 0.5, \lambda_e = 0.3$时，模型取得最佳综合性能，其平均F1值和平均IoU分别为69.5%和61.1%。上述结果表明，权重系数的设置需要同时兼顾篡改主体区域建模与边界细节约束。基于上述实验结果与损失设计动机，后续实验中的权重系数设置为$\lambda_b = 0.5, \lambda_e = 0.3$。

3.3 对比实验结果分析

为验证所提方法的有效性与泛化能力，本文选取并复现了13种具有代表性的图像篡改检测与定位方法作为对比模型。具体包括空—频域融合方法H-LSTM[10]和CAT-Net v2[26]，经典空域取证方AdaCFA[27]，基于Vision Transformer的Swin-ViT[28]与EVP[29]，多尺度与跨域检测方法MVSS-Net++[30]、CFLNet[31]、EITLNet[32]、PIM[3]和FakeShield[33]，基于稀疏注意力建模的Sparse-ViT[4]、基于局部–全局中继注意力机制的Relay-Former[34]和基于噪声感知交叉注意力的NC-Net[35]等方法。

为保证对比实验的公平性，本文统一各方法的测试集划分与评价指标，并尽可能按照其原始训练数据配置进行复现。对于公开训练代码的模型，依据论文或官方实现所采用的数据配置重新训练。对于未提供完整训练代码或训练策略的模型，则采用其官方预训练模型进行测试，并保持原始输入预处理方式不变。在此基础上，通过统一测试协议对各方法进行评估，以减小实现差异对结果的影响。

对比实验在CASIA v1、MISD、Coverage、Columbia、NIST16以及CocoGlide六个公开基准数

表3　不同损失权重系数组合下的篡改定位性能比较[%]

λ_b	λ_e	CASIA v1		MISD		Coverage		Columbia		NIST16		CocoGlide		AVG	
		F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU
0.3	0.1	81.7	75.8	71.6	58.4	61.7	53.7	80.3	75.0	47.0	40.9	61.8	52.6	67.4	59.4
0.3	0.3	82.8	76.7	69.3	56.0	61.2	53.4	**82.8**	**77.5**	50.8	44.6	62.0	52.8	68.2	60.2
0.3	0.5	82.9	76.3	64.4	50.8	60.8	52.4	79.0	72.9	50.1	43.3	59.0	48.9	66.0	57.4
0.5	0.1	80.1	73.4	68.4	55.1	56.4	48.6	78.8	72.7	46.7	40.3	58.6	48.7	64.8	56.5
0.5	**0.3**	82.6	75.9	**72.0**	**59.0**	**61.9**	**53.9**	81.4	75.6	**53.2**	**46.4**	**65.7**	**56.0**	**69.5**	**61.1**
0.5	0.5	80.7	74.2	65.0	51.3	55.7	48.0	79.5	73.7	45.8	40.0	55.5	45.9	63.7	55.5
0.7	0.1	**83.7**	**77.7**	71.0	57.6	60.0	52.8	81.8	76.5	50.8	44.8	64.6	54.7	68.7	60.7
0.7	0.3	79.9	73.6	67.5	53.9	54.2	47.8	78.9	72.8	44.5	38.5	62.1	52.8	64.5	56.6
0.7	0.5	79.6	72.8	67.1	53.2	60.2	53.4	79.6	73.5	44.2	38.3	62.2	52.1	65.5	57.2

据集上进行，并从像素级定位与图像级检测两个层面进行评估。其中，F1 值在固定阈值 0.5 下计算。表中加粗数值表示最优结果，下划线表示次优结果。对于跨数据集综合性能，本文采用各测试数据集评价结果的算术平均（average, AVG）进行统计，而非基于样本数量的加权平均。该统计方式能够减弱大规模数据集对综合指标的主导作用，更侧重反映模型在不同数据分布下的整体泛化能力。因此，AVG 指标用于描述模型的跨数据集平均表现，而不用于严格的统计显著性推断。

3.3.1　像素级定位性能分析

像素级篡改定位结果如表 4 所示。不同方法在各数据集上的性能差异较为明显，表明了篡改定位对篡改类型及数据分布具有较强敏感性。部分方法在特定场景下表现突出，但在跨数据集场景下往往存在显著性能波动。Swin-ViT 在 MISD 数据集上的 F1 为 70.2%，但在 Coverage 数据集上仅为 10.4%。SparseViT 在 Columbia 数据集上的 F1 达到 95.2%，但在 NIST16 数据集上仅为 30.7%。结果表明，依赖单一建模机制的方法虽可在特定篡改类型上获得较高精度，但在面对篡改模式与数据分布变化时，整体泛化能力有限。

相比之下，本文方法在多个数据集上表现出更好的稳定性，在 CASIA v1、NIST16 和 CocoGlide 数据集上均取得最优结果，多数据集平均 F1 值和平均 IoU 分别达到 69.5% 和 61.1%。在 MISD、Coverage 和 Columbia 数据集上，本文方法的 F1 值分别为 72.0%、61.9% 和 81.4%，略低于 NC-Net、RelayFormer 和 SparseViT 的最优结果。该差异主要源于建模侧重点的不同。SparseViT 与 RelayFormer 基于 Transformer 的自注意力机制能够更有效地刻画长距离依赖与结构关系，NC-Net 通过交叉注意力增强对局部细粒度异常的响应能力。相比之下，本文方法通过语义特征与取证残差的协同建模，更侧重语义一致性与统计异常的联合约束，在边界不连续与统计扰动建模方面具有优势，但对显式结构关系的刻画能力相对有限，因此在结构主导的篡改场景中表现略低。综上所述，本文方法能够在不同篡改类型与数据分布条件下兼顾区域检测能力与边界刻画精度，从而获得更稳定的定位性能。

3.3.2　图像级检测性能分析

图像级检测结果如表 5 所示。传统统计特征方法 H-LSTM 与 AdaCFA 整体性能较低，其平均 AUC 分别仅为 53.3% 和 54.6%，平均 EER 分别为 47.6% 和 46.9%，接近随机判别水平，说明仅依赖局部统计特征或浅层取证线索难以有效区分复杂篡改图像与真实图像。基于深度特征建模的方法整体性能有所提升，例如 CAT-Net v2 的平均 AUC 达到 79.8%，FakeShield 在 CocoGlide 数据集上的 AUC 达到 82.8%。然而，这类方法对数据分布变化和篡改模式迁移的适应能力仍然有限。相比之下，本文方法在各测试数据集上表现出更好的整体稳定性。多数据集平均 AUC 达到 90.7%，较 SparseViT 的 84.0% 提高了 6.7%。平均 EER 低至 12.7%，较 SparseViT 的 20.9% 降低 8.2%。结果表明，本文方法在不同篡改类型与数据分布条件下均具有较好的检测稳定性。

综合像素级定位与图像级检测结果可以看出，

表 4 图像伪造定位的像素级 F1 和 IoU 性能表现[%]

方法	CASIA v1		MISD		Coverage		Columbia		NIST16		CocoGlide		AVG	
	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU
H-LSTM [10]	15.5	10.0	29.4	17.3	16.8	10.6	13.8	8.1	34.2	26.1	15.9	10.0	20.9	13.7
AdaCFA [27]	16.3	8.7	29.0	17.9	20.4	11.7	39.5	29.0	11.3	6.5	31.1	19.7	24.6	15.6
Swin-ViT [28]	40.7	36.4	70.2	57.4	10.4	8.1	26.1	21.1	21.7	17.4	9.7	7.5	29.8	24.7
CAT-Net v2 [26]	71.0	63.6	40.1	31.8	28.5	22.7	79.9	74.9	16.4	13.3	36.4	28.8	45.4	39.2
MVSS-Net++ [30]	51.2	44.4	68.0	54.9	48.1	41.2	69.1	60.3	30.4	24.5	44.1	34.6	51.8	43.3
CFLNet [31]	16.3	11.0	33.6	22.3	15.9	10.3	22.8	15.6	18.3	12.6	17.9	11.7	20.8	13.9
EVP [29]	47.9	41.7	59.8	46.5	10.3	7.4	33.0	25.4	23.9	18.8	12.1	8.7	31.2	24.8
EITLNet [32]	55.7	52.0	74.7	62.9	42.1	33.4	78.6	75.8	34.3	28.3	35.4	28.8	53.5	46.9
PIM [3]	57.0	51.7	73.4	60.5	26.7	20.6	66.0	58.4	29.3	23.7	34.0	27.0	47.7	40.3
FakeShield [33]	56.3	50.9	51.5	37.5	21.3	18.8	72.0	63.7	23.2	20.6	53.0	43.4	46.2	39.2
SparseViT [4]	81.1	75.9	76.5	64.4	55.6	51.3	**95.2**	**93.6**	30.7	26.0	39.7	33.4	63.1	57.5
RelayFormer [34]	80.8	73.9	66.6	53.7	**63.1**	**57.4**	87.7	85.8	37.8	34.3	30.7	24.5	61.1	54.9
NC-Net [35]	81.7	75.4	**77.7**	**65.7**	61.7	53.4	93.3	90.4	36.8	29.1	54.6	45.4	67.6	59.9
本文方法	**82.6**	**75.9**	72.0	59.0	61.9	53.9	81.4	75.6	**53.2**	**46.4**	**65.7**	**56.0**	**69.5**	**61.1**

现有方法在不同数据集上的性能差异具有明显的数据依赖性。部分模型虽能在特定篡改模式下取得较优结果，但在跨数据集或跨篡改类型场景中往往出现明显性能波动，反映出单一特征建模方式难以同时兼顾语义信息与取证线索，从而限制了模型的泛化能力。从建模机制上看，基于 Transformer 的方法在结构关系显著的场景中更具优势，但在复杂退化或非结构性伪影主导的条件下稳定性不足。而依赖局部统计或频域特征的方法虽对特定噪声模式敏感，但对多样化篡改形式的适应能力有限。相比之下，本文方法通过语义特征与取证特征的协同建模，并结合两阶段训练策略与置信度感知机制，增强篡改区域表征能力与图像级判别能力，从而提升模型在复杂数据分布条件下的鲁棒性与泛化能力。

3.4 消融实验结果分析

为进一步分析各组成模块对模型性能的贡献，本文在 CASIA v1、Coverage、NIST16 和 CocoGlide 四个具有代表性的数据集上开展消融实验。实验分别从特征编码模块、取证特征提取方式、特征融合策略、检测头特征组分以及任务解耦机制五个方面进行分析。通过构建逐步替换与累积组合的对照实验设置，对各模块的独立贡献及其协同作用进行系统评估，从而验证整体模型设计的有效性和各模块配置的合理性。

3.4.1 特征编码模块消融

特征编码模块的消融结果如表 6 所示。仅采用 DinoV2 ViT-B/14 作为语义编码器时，模型整体定位性能较低，说明单一语义表征难以刻画篡改区域的细粒度异常。引入取证分支构建语义—取证双流结构后，各数据集性能均有所提升，CASIA v1 的 F1 由 51.7% 提升至 58.3%，IoU 由 41.9% 提升至 49.4%，表明低层取证特征能够补充语义特征在局部统计异常与边界扰动建模上的不足。进一步将主干由 ViT-B 扩展至 ViT-L 后，不同数据集表现不一致，CASIA v1 数据集上的 F1 由 58.3% 提升至 59.0%，而 Coverage 和 NIST16 数据集上的 F1 分别下降至 26.1% 和 37.0%，对应 IoU 亦出现下降，表明单纯依赖主干规模难以稳定提升跨场景泛化能力。在此基础上引入小波域注意力模块与多尺度 FPN 后，各数据集的 F1 和 IoU 均显著提升，CASIA v1、Coverage、NIST16 和 CocoGlide 数据集上的 F1 分别提升至 82.6%、61.9%、53.2% 和 65.7%。该结果表明，小波域高频增强与多尺度建模能够有效提升对复杂篡改区域及边界细节的表达能力。

3.4.2 取证特征提取方式消融

取证特征提取方式的消融结果如表 7 所示。当仅采用单一取证提取器时，不同方法在不同数据集上表现存在差异。SRM 在 CASIA v1 和 CocoGlide

表 5　图像伪造检测的图像级 EER 和 AUC 性能表现[%]

方法	CASIA v1		MISD		Coverage		Columbia		NIST16		CocoGlide		AVG	
	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC
H-LSTM [10]	50.3	48.2	47.6	53.3	49.0	53.7	42.2	61.4	45.9	55.0	50.4	48.2	47.6	53.3
AdaCFA [27]	46.2	54.4	42.5	60.1	48.7	52.4	42.7	62.0	49.8	50.4	51.7	48.3	46.9	54.6
Swin-ViT [28]	28.8	79.9	3.0	99.5	50.0	50.4	26.1	81.4	44.3	59.7	45.9	53.8	33.0	70.8
CAT-Net v2 [26]	13.4	94.2	20.3	88.9	37.0	68.3	6.7	97.7	42.5	62.8	38.3	66.7	26.4	79.8
MVSS-Net++ [30]	22.4	85.7	2.4	99.3	34.0	69.6	5.6	97.8	43.9	58.6	39.3	66.4	24.6	79.6
CFLNet [31]	48.5	50.2	38.5	65.9	49.0	50.5	31.7	74.6	47.1	53.7	50.2	50.8	44.2	57.6
EVP [29]	25.0	82.3	7.8	96.3	47.0	53.6	35.0	70.2	41.1	63.2	47.1	53.7	33.8	69.9
EITLNet [32]	22.9	84.5	3.8	99.3	41.2	61.9	18.2	91.0	39.8	61.2	38.0	64.7	27.3	77.1
PIM [3]	31.2	76.6	**1.9**	**99.9**	45.0	56.5	16.7	89.6	42.9	60.6	41.2	62.6	29.8	74.3
FakeShield [33]	17.4	90.6	10.5	95.2	47.0	56.9	8.3	96.5	32.3	74.6	22.1	82.8	22.9	82.8
SparseViT [4]	13.0	93.7	3.6	99.4	28.0	78.9	3.3	99.7	39.8	65.4	37.9	67.0	20.9	84.0
RelayFormer [34]	12.3	93.8	10.5	95.7	**24.0**	**80.5**	2.8	99.7	46.5	56.0	38.1	66.0	22.4	82.0
NC-Net [35]	18.0	90.0	4.7	97.4	32.0	74.7	5.0	98.7	39.3	63.0	32.6	72.9	21.9	82.8
本文方法	**8.7**	**95.9**	2.0	99.7	25.0	78.8	**1.1**	**99.9**	**24.7**	**81.7**	**14.5**	**88.1**	**12.7**	**90.7**

数据集上的 F1 分别为 82.0% 和 58.8%，而 Bayar 在 Coverage 和 NIST16 数据集上表现更优，F1 分别达到 57.2% 和 47.6%，对应 IoU 亦呈现相似趋势。从特征建模机制来看，SRM 通过预定义高通滤波器提取固定残差信息，对纹理扰动和高频异常具有较稳定的响应，但对数据分布变化的自适应能力有限。Bayar 约束卷积通过可学习滤波核生成残差特征，能够根据数据特性动态调整，对局部统计约束破坏及边界不连续性更为敏感，但其表达受训练数据分布影响较大。因此，单一取证特征难以同时兼顾不同类型篡改伪影。

在此基础上，将 SRM 与 Bayar 结合后，各数据集性能均优于仅使用单一取证特征的结果，说明固定残差建模与自适应残差建模具有明显互补性。两类特征在频域响应与结构约束上的协同作用，使模型能够同时刻画纹理异常与边界不连续性，从而提升在不同数据分布下的稳定性。

3.4.3　特征融合模块消融

特征融合模块的消融结果如表 8 所示。在基线融合方式上引入 FA 后，模型在不同数据集上的表现存在明显差异。在 CASIA v1 和 NIST16 数据集上性能有所提升，F1/IoU 分别达到 82.9%/76.6% 和 50.1%/43.6%，而 Coverage 和 CocoGlide 数据集上则呈现下降趋势。这一现象表明，特征对齐能够在一定程度上缓解跨分支特征分布偏移，但其对不同数据分布的适应性有限。当篡改模式复杂或跨域差异较大时，强制对齐可能削弱原始特征中的判别信息，从而导致性能下降。在此基础上进一步引

表 6　特征编码模块的消融实验结果[%]

版本	CASIA v1		Coverage		NIST16		CocoGlide	
	F1	IoU	F1	IoU	F1	IoU	F1	IoU
DinoV2 ViT-B/14	51.7	41.9	20.7	14.2	33.6	25.6	43.6	33.8
DinoV2 ViT-B/14+forensic	58.3	49.4	26.7	18.4	40.3	32.2	46.6	36.6
DinoV2 ViT-L/14+forensic	59.0	51.1	26.1	20.1	37.0	30.5	45.6	36.7
DinoV2 ViT-L/14+forensic+wavelet	65.8	57.6	38.4	29.8	43.0	36.0	52.0	42.0
DinoV2 ViT-L/14+forensic+wavelet+FPN	**82.6**	**75.9**	**61.9**	**53.9**	**53.2**	**46.4**	**65.7**	**56.0**

表7　取证特征提取方式的消融实验结果[%]

版本	CASIA v1		Coverage		NIST16		CocoGlide	
	F1	IoU	F1	IoU	F1	IoU	F1	IoU
Only_SRM	82.0	**76.1**	56.7	50.6	39.2	34.7	58.8	50.0
Only_Bayar	81.9	75.8	57.2	49.4	47.6	41.3	52.8	43.6
SRM+Bayar	**82.6**	75.9	**61.9**	**53.9**	**53.2**	**46.4**	**65.7**	**56.0**

入 GF 后，四个数据集上的性能均出现下降，其中 NIST16 数据集下降最为明显，F1 和 IoU 由 50.1% 和 43.6% 下降至 41.7% 和 35.2%。说明在缺乏有效约束的情况下，GF 在多源特征动态加权过程中容易引入不稳定的特征选择偏差，从而削弱语义与取证信息的协同表达能力。加入 ER 后，模型性能在各数据集上均得到恢复并提升，尤其在 Coverage、NIST16 和 CocoGlide 等复杂场景下提升更为明显。这说明边缘细化过程能够增强对篡改区域边界与局部结构信息的约束，有效弥补融合过程中细节信息的损失。综上，融合性能的提升并非来源于单一模块的简单叠加，而依赖于特征对齐、动态交互与边界增强之间的协同作用，从而实现语义信息与取证特征的稳定融合。

3.4.4　检测头特征消融

检测头特征组合的消融结果如表 9 所示。在基线检测头中引入语义特征 F_{dino} 后，模型在各数据集上的检测性能明显提升，在 CASIA v1 数据集上的 AUC 由 47.0% 提升至 75.2%，EER 由 52.2% 降至 31.4%，说明高层语义表征能够提供有效的全局判别信息。在此基础上加入取证特征 F_{sap} 后，模型性能显著提升，CASIA v1、Coverage、NIST16 和 CocoGlide 数据集上的 AUC 分别达到 95.7%、59.9%、81.8% 和 71.5%。该结果表明，语义信息与取证特征的联合建模能够有效增强对伪造痕迹的判别能力。引入频域特征 F_{wsp} 后，模型在 CASIA v1 和 NIST16 数据集上仍有小幅提升，但在 Coverage 和 CocoGlide 数据集上出现明显下降，这表明频域特征对数据分布较为敏感，在复杂或跨域场景中可能引入冗余信息，从而干扰判别过程。加入多尺度特征 F_{mmp} 后，模型性能得到恢复并取得最优结果，Coverage 和 CocoGlide 数据集上的 AUC 分别提升至 78.8% 和 88.1%。说明多尺度信息能够缓解单一特征带来的不稳定性，通过跨尺度建模提升检测鲁棒性。综上，检测性能的提升并非依赖单一特征，而是来源于多类特征组分之间的互补建模。

3.4.5　不同任务解耦方式消融

不同任务解耦方式对定位与检测性能的影响如表 10 所示。定位与检测任务虽具有不同优化目标，但在底层伪造线索上存在共享基础，因此解耦方式会对两类任务产生不同影响。对于定位任务，无解耦在 CASIA v1、Coverage 和 NIST16 数据集上的 F1 分别为 70.2%、40.3% 和 31.5%，均高于硬解耦。说明完全分离特征表示会削弱定位分支对共享取证线索的利用能力，从而影响篡改区域的精细刻画。在 CocoGlide 数据集上，硬解耦的 F1 由 52.6%

小幅提升至 53.4%，表明在生成式篡改场景中，适当降低任务间干扰有助于缓解语义与取证信息的不一致问题。对于检测任务，硬解耦仅在 CASIA v1 数据集上带来轻微提升，在 Coverage、NIST16 和 CocoGlide 数据集上分别下降至 75.3%、74.2% 和 93.0%。表明完全解耦削弱了跨任务共享信息对图像级判别的支撑作用，导致性能不稳定。

相比之下，软解耦在定位与检测两个任务上均取得更优结果。其通过在共享特征基础上引入轻量化任务适配，使模型在保留共享伪造线索的同时，对任务差异进行有效调制，从而在信息共享与任务特异性之间取得更合理的平衡。综上，任务解耦方式本质上体现为共享与独立建模之间的权衡。软解耦能够在保证信息共享的同时提升任务适配能力，从而获得更稳定的综合性能。

表8　特征融合模块的消融实验结果[%]

版本	CASIA v1		Coverage		NIST16		CocoGlide	
	F1	IoU	F1	IoU	F1	IoU	F1	IoU
Baseline	81.6	74.7	**66.8**	**57.9**	50.0	43.3	64.6	54.3
+FA	**82.9**	**76.6**	62.0	54.0	50.1	43.6	61.4	52.2
+FA+GF	78.4	71.4	56.9	49.9	41.7	35.2	56.5	46.7
+FA+GF+ER	82.6	75.9	61.9	53.9	**53.2**	**46.4**	**65.7**	**56.0**

表 9　检测头特征组合的消融实验结果[%]

版本	CASIA v1		Coverage		NIST16		CocoGlide	
	EER	AUC	EER	AUC	EER	AUC	EER	AUC
Baseline	52.2	47.0	54.0	43.9	43.4	61.2	45.5	55.9
$+F_{dino}$	31.4	75.2	47.0	55.1	42.5	61.3	39.1	66.8
$+F_{dino}+F_{sap}$	9.5	95.7	43.0	59.9	27.1	81.8	34.8	71.5
$+F_{dino}+F_{sap}+F_{wsp}$	9.4	95.8	56.0	42.4	24.7	**82.1**	59.6	43.3
$+F_{dino}+F_{sap}+F_{wsp}+F_{mmp}$	**8.7**	**95.9**	**25.0**	**78.8**	**24.7**	81.7	**14.5**	**88.1**

表 10　不同任务解耦方式对定位与检测性能的影响[%]

版本	CASIA v1		Coverage		NIST16		CocoGlide	
	定位(F1)	检测(AUC)	定位(F1)	检测(AUC)	定位(F1)	检测(AUC)	定位(F1)	检测(AUC)
无解耦	70.2	94.2	40.3	77.0	31.5	81.5	52.6	**93.3**
硬解耦	65.8	94.7	23.9	75.3	19.8	74.2	53.4	93.0
软解耦	**82.6**	**95.9**	**61.9**	**78.8**	**53.2**	**81.7**	**65.7**	88.1

3.5　可视化分析

为直观评估模型在复杂篡改场景下的定位能力与预测可靠性，本文在 CASIA v1 测试集中选取具有代表性的篡改样本进行可视化分析。图 4 给出了不同方法在典型复杂样本上的定位结果对比，包括 EVP、EITLNet、PIM、FakeShield 以及本文方法。图 5 展示了本文方法的定位结果及对应的像素级置信度分布。其中，预测掩码表示模型输出的定位响应图，置信度图表示置信度分支输出的像素级可靠性估计，灰度越亮表示置信度越高，反之则表示不确定性更强。

如图 4 所示，不同方法在复杂场景下呈现出明显

差异。EVP 在小尺度篡改区域上响应较弱，易出现漏检。EITLNet 在复杂背景区域产生明显伪响应。PIM 在目标边界处存在扩张与断裂现象，边界刻画不够精确。FakeShield 虽能捕捉主体区域，但在细粒度结构处响应不连续。相比之下，本文方法在多数样本上能够生成更完整、连续的篡改区域响应，预测边界与真实标注具有更高一致性，表明语义特征与取证特征的协同建模有助于同时刻画区域级语义约束与局部异常线索，从而提升复杂场景下的定位稳定性。

图 5 进一步展示了本文方法的置信度分布特性。可以观察到，高置信度区域主要集中在篡改主体内部，低置信度响应则多分布于目标边界、拼接接缝及纹理复杂区域。该分布与篡改区域的实际不确定性较为一致，说明所学习到的置信度表征能够较好反映定位结果的可靠性差异，而非简单重复定位响应。综合图 4 与图 5 可以看出，本文方法不仅能够在复杂场景下获得更准确的篡改区域定位结果，还能够对预测可靠性进行有效表征。这为由像素级定位向图像级判别的特征聚合提供了更稳健的依据，也进一步验证了置信度感知机制的有效性。

3.6　鲁棒性分析

为评估模型在图像质量退化条件下的稳定性，本文在 CASIA v1 和 Columbia 数据集上开展鲁棒性实验，结果如图 6 和图 7 所示。实验模拟真实传输与存储过程中的常见退化形式，JPEG 压缩的质量因子 Q 从 100 逐步降低至 50，高斯噪声的标准差 σ 从 0 增加至 25，高斯模糊的卷积核尺寸 k 从 1 扩展至 9。评价指标采用图像级 AUC 和像素级 F1。

从整体趋势来看，随着退化强度增加，各方法性能均呈下降趋势，但不同模型的下降幅度存在明显差异。对于 JPEG 压缩，PIM 与 EITLNet 的 AUC 和 F1 随压缩增强持续下降，在低质量区间衰减更为明显，F1 和 AUC 的下降幅度约在 20% 左右。相比之下，本文方法的性能曲线变化更为平缓，说明其对压缩引起的高频信息损失具有更强适应性。在高斯噪声条件下，对比方法的性能随噪声增强快速下降，尤其在 F1 指标上更为明显，表明随机扰动更容易破坏其所依赖的局部纹理与边界信息。而本

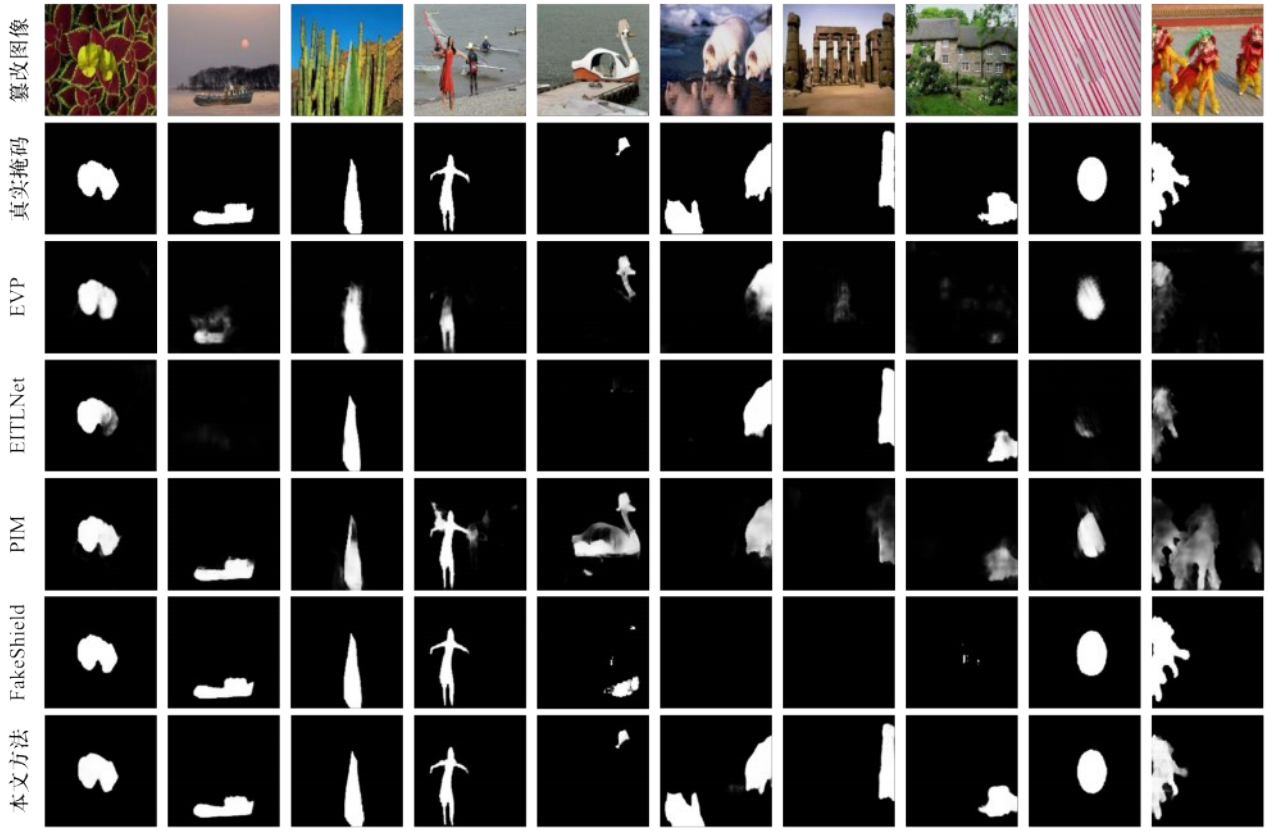


图4　不同方法在典型篡改样本上的定位结果对比

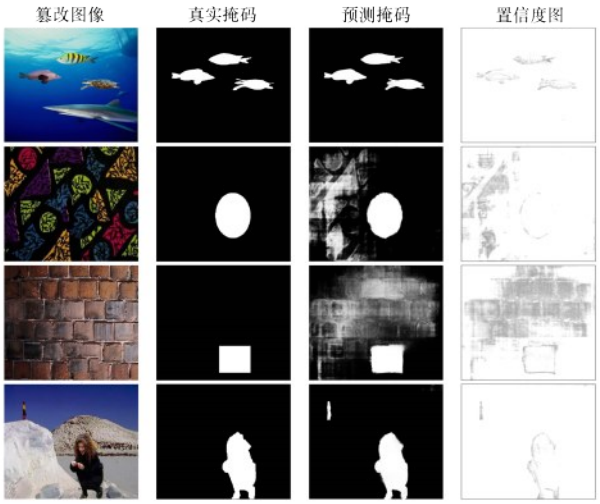


图5　本文方法的篡改定位结果及像素级置信度分布

文方法在噪声增强过程中仍能保持较稳定的 AUC 和 F1，体现出更强的抗噪判别能力。对于高斯模糊，各方法性能下降相对缓和，但对比方法仍表现出持续衰减趋势，而本文方法的下降幅度最小，表明其在边界扩散和局部结构平滑条件下仍能保留较有效的判别信息。进一步观察可知，AUC 的波动整体小于 F1，说明图像级判别对质量退化相对不敏感，而像素级定位由于更依赖高频纹理和边界细节，在退化增强时更易受到影响。

综合来看，本文方法在三类退化条件下均表现出较低的性能衰减率，表现出较好的鲁棒性。这表明，语义信息与取证线索的协同建模以及多尺度特征融合，有助于模型在图像质量受损时保持较稳定的结构表征与判别能力，从而体现出更好的鲁棒性。

4　结束语

针对深度伪造取证任务中异构特征冲突、高频线索衰减以及多任务训练带来的负迁移问题，本文提出一种基于多特征协同与置信度感知的深度伪造图像检测与定位方法。该方法通过语义—取证双流结构与小波增强模块强化微弱篡改痕迹的表达，并结合边缘细化机制提升对细粒度篡改区域的定位能力。在图像级检测任务中，通过引入任务适配器与置信度校准模块构建独立检测分支，实现定位与检测任务的软解耦。实验结果表明，所提方法在多个公开数据集上均取得较优性能，在像素级定位与图像级检测任务中均表现出良好的泛化性与鲁棒性。

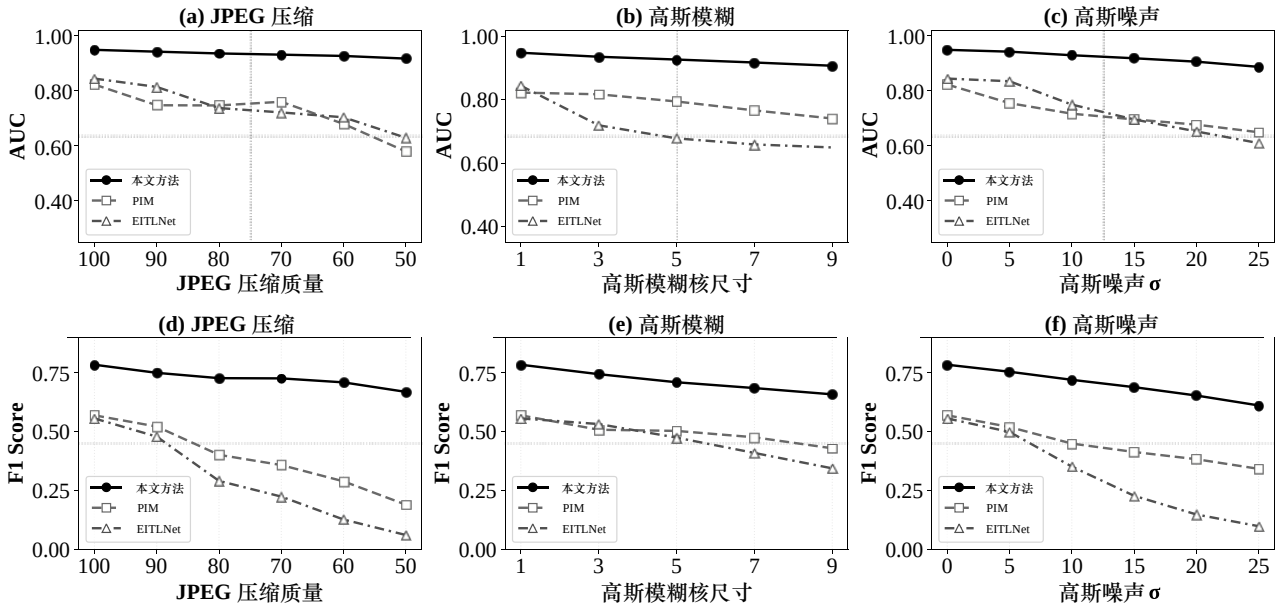


图6　CASIA v1 数据集上图像级 AUC 和像素级 F1 的鲁棒性对比实验结果

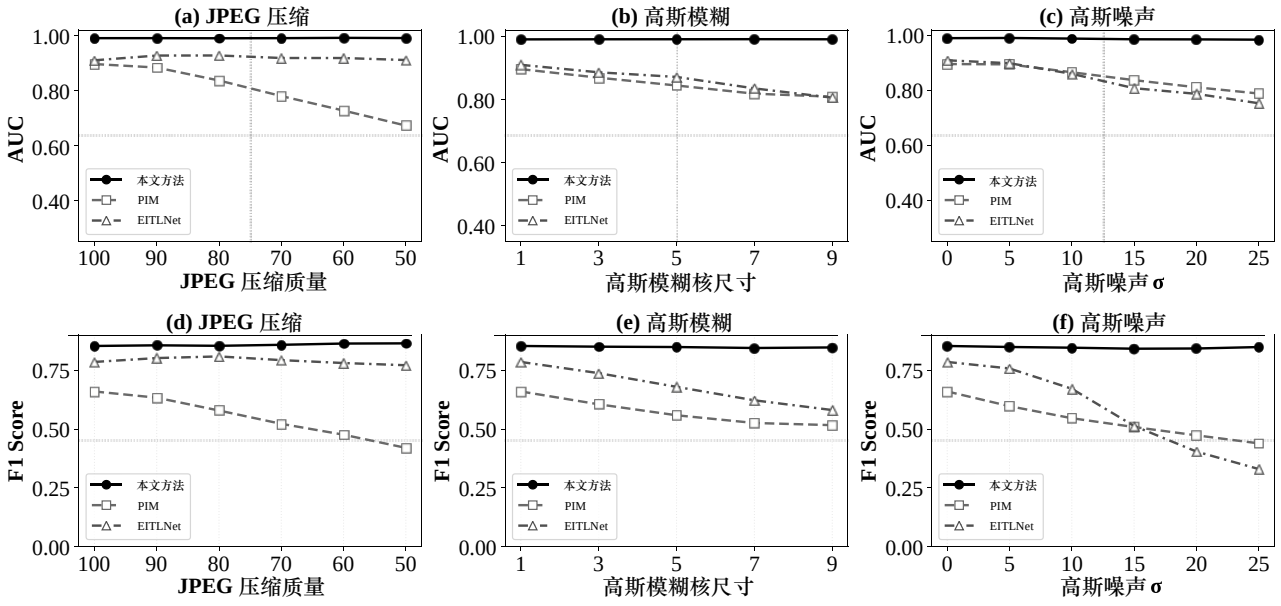


图7　Columbia 数据集上图像级 AUC 和像素级 F1 的鲁棒性对比实验结果

然而，本文模型依赖较复杂的特征协同结构，在高分辨率图像场景下计算开销较高。当前训练数据主要来源于已知篡改类型，在面对新型生成模型或未知编辑范式时仍可能存在性能下降。未来将探索知识蒸馏与动态推理的轻量化部署方案，进一步挖掘扩散模型逆向过程中的频域不变特征，以提升开放场景下对未知生成范式的零样本迁移能力。

参考文献：

[1] Kadha v k, 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.

[2] Pei g, Zhang j, Hu m, et al. Deepfake generation and detection: a benchmark and survey[J]. ACM Computing Surveys, 2026, 58(11): 273: 1-273:41.

[3] 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.

[4] Su l, Ma x, Zhu x, et al. Can we get rid of handcrafted feature extrac-

tors? SparseViT: nonsemantics-centered, parameter-efficient image manipulation localization through spare-coding transformer[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia, PA, USA, 2025-02-25/2025-03-04. 2025, 39(7): 7024-7032.

[5] Guillaro f, Cozzolino d, Sud a, et al. TruFor: leveraging all-round clues for trustworthy image forgery detection and localization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, BC, Canada, 2023-06-18/2023-06-22: 20606-20615.

[6] Li y, Zhou j. Fast and effective image copy-move forgery detection via hierarchical feature point matching[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(5): 1307-1322.

[7] 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.

[8] Matern f, Riess c, Stamminger m. Gradient-based illumination description for image forgery detection[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 1303-1317.

[9] Bayar b, Stamm m c. Constrained convolutional neural networks: a new approach towards general purpose image manipulation detection[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(11): 2691-2706.

[10] 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.

[11] Hao j, Zhang z, Yang s, et al. TransForensics: image forgery localization with dense self-attention[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, QC, Canada, 2021-10-10/2021-10-17: 15055-15064.

[12] Wang j, Wu z, Chen j, et al. ObjectFormer for image manipulation detection and localization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA, 2022-06-19/2022-06-24: 2364-2373.

[13] Qian y, Yin g, Sheng l, et al. Thinking in frequency: face forgery detection by mining frequency-aware clues[C]//Computer Vision – ECCV 2020: 16th European Conference. Glasgow, UK, 2020-08-23/2020-08-28. Cham: Springer, 2020: 86-103.

[14] Liu h, Li x, Zhou w, et al. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021-06-19/2021-06-25: 772-781.

[15] Luo y, Zhang y, Yan j, et al. Generalizing face forgery detection with high-frequency features[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021-06-19/2021-06-25: 16317-16326.

[16] 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. Vancouver, BC, Canada, 2024-02-20/2024-02-27. 2024, 38(5): 5052-5060.

[17] Li h, Zhou j, Li y, et al. FreqBlender: enhancing deepfake detection by blending frequency knowledge[J]. Advances in Neural Information Processing Systems, 2024, 37: 44965-44988.

[18] Huang z, Hu j, Li x, et al. SIDA: social media image deepfake detection, localization and explanation with large multimodal model[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA, 2025-06-11/2025-06-15: 28831-28841.

[19] Yu t, Kumar s, Gupta a, et al. Gradient surgery for multi-task learning [J]. Advances in Neural Information Processing Systems, 2020, 33: 5824-5836.

[20] Shi h, Ren s, Zhang t, et al. Deep multitask learning with progressive parameter sharing[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023-10-01/2023-10-06: 19924-19935.

[21] Peng s, Zhang t, Gao l, et al. WMamba: wavelet-based mamba for face forgery detection[C]//Proceedings of the 33rd ACM International Conference on Multimedia. Dublin, Ireland, 2025-10-27/2025-10-31: 4768-4777.

[22] Song y, Xu y, Chen j, et al. Wavelet convolution and multi-scale attention network for image tampering localization[C]//2025 IEEE International Conference on Multimedia and Expo. Nantes, France, 2025-06-30/2025-07-04: 1-6.

[23] Kniaz v v, Knyaz v a, Remondino f. The point where reality meets fantasy: mixed adversarial generators for image splice detection[C]//Advances in Neural Information Processing Systems 32. Vancouver, Canada, 2019-12-08/2019-12-14: 215-226.

[24] Novozamsky a, Mahdian b, Saic s. IMD2020: a large-scale annotated dataset tailored for detecting manipulated images[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops. Snowmass Village, CO, USA, 2020-03-01/2020-03-05: 71-80.

[25] Dong j, Wang w, Tan t. CASIA image tampering detection evaluation database[C]//2013 IEEE China Summit and International Conference on Signal and Information Processing. Beijing, China, 2013-07-06/2013-07-10: 422-426.

[26] 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.

[27] Bammey q, Gioi r g v, Morel j m. An adaptive neural network for unsupervised mosaic consistency analysis in image forensics[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2020-06-14/2020-06-19: 14182-14192.

[28] Liu z, Lin y, Cao y, et al. Swin Transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, QC, Canada, 2021-10-11/2021-10-17: 10012-10022.

[29] Liu w, Shen x, Pun c m, et al. Explicit visual prompting for low-level structure segmentations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, BC, Canada, 2023-06-18/2023-06-22: 19434-19445.

[30] 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.

[31] Niloy f f, Bhaumik k k, Woo s s. CFL-Net: image forgery localization using contrastive learning[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, HI, USA, 2023-01-03/2023-01-07: 4631-4640.

[32] Guo k, Zhu h, Cao g. Effective image tampering localization via enhanced transformer and co-attention fusion[C]//2024 IEEE Interna-

tional Conference on Acoustics, Speech and Signal Processing. Seoul, Korea, 2024-04-14/2024-04-19: 4895-4899.

[33] Xu z, Zhang x, Li r, et al. FakeShield: explainable image forgery detection and localization via multi-modal large language models[C]//The Thirteenth International Conference on Learning Representations. Singapore, 2025-04-24/2025-04-28.

[34] Huang w, Yang j, Dai t, et al. RelayFormer: a unified local-global attention framework for scalable image and video manipulation localization [C]//The Fourteenth International Conference on Learning Representations. Rio de Janeiro, Brazil, 2026-04-23/2026-04-27.

[35] Zhang h, Su t, Liu z, et al. Noise-aware cross attention for image manipulation localization[J]. Pattern Recognition, 2026, 176: 113164.

[36] Kadam k d, Ahirrao s, Kotecha k. Multiple image splicing dataset (MISD): a dataset for multiple splicing[J]. Data, 2021, 6(10): 102.

[37] Wen b, Zhu y, Subramanian r, et al. COVERAGE: a novel database for copy-move forgery detection[C]//2016 IEEE International Conference on Image Processing. Phoenix, AZ, USA, 2016-09-25/2016-09-28: 161-165.

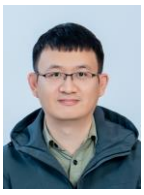
[38] Hsu y f, Chang s f. Detecting image splicing using geometry invariants and camera characteristics consistency[C]//2006 IEEE International Conference on Multimedia and Expo. Toronto, ON, Canada, 2006-07-09/2006-07-12: 549-552.

[39] Guan h, Kozak m, Robertson e, et al. MFC datasets: large-scale benchmark datasets for media forensic challenge evaluation[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops. Waikoloa, HI, USA, 2019-01-07/2019-01-11: 63-72.

郭祯（1981− ），女，博士，海南大学副教授，主要研究方向为数据安全、隐私保护与区块链技术。

邱润尧（2002− ），男，海南大学硕士生，主要研究方向为网络与信息安全、深度伪造检测。

徐嘉（1997− ），女，博士，海南大学博士生，主要研究方向为医学图像处理、数字水印、深度神经网络和计算机视觉等。



刘志全（1989− ），男，博士，暨南大学教授，主要研究方向为车联网安全、低空安全、Web 安全、信任建模、隐私计算、区块链、人工智能等。

马建峰（1963− ），男，博士，西安电子科技大学教授，主要研究方向为网络安全、系统安全、数据安全和无人机安全等。