基于多粒度表征解耦与协同的高泛化图像篡改检测与定位方法

袁程胜[1,2]，曹富贵[1,2]，王一力[1,2]，曹燚[3]，刘庆程[1]，付章杰[1,2]
（1.南京信息工程大学网络空间安全学院，江苏 南京 210044；2.南京信息工程大学数字取证教育部工程研究中心，江苏 南京 210044；
3.无锡学院物联网工程学院，江苏 无锡 214105）

摘要：移动宽带传输图像等多媒体内容时易遭受截获与篡改，威胁通信内容的真实性。针对现有方法跨数据集泛化能力不足及复杂边界、小区域篡改定位不准的问题，提出一种基于多粒度表征解耦与协同的高泛化图像篡改检测与定位方法。首先，构建细粒度伪迹表征与粗粒度结构表征，分别刻画局部伪迹和区域结构信息。其次，引入显式解耦约束，降低不同粒度表征间的冗余相关性。最后，设计伪迹先验引导的门控自适应融合模块，动态调节局部细节与上下文结构信息的贡献。实验结果表明，所提方法在CASIA_V1、Columbia、NIST16和Coverage等公开数据集的跨数据集测试中取得较好综合性能，具有较强的泛化能力和定位精度。

关键词：移动宽带；图像篡改检测与定位；多粒度表征；显式解耦；门控自适应融合
中图分类号：TP391
文献标志码：A
doi：10.11959/j.issn.1000–436x.TXXB260219

Highly generalizable image tampering detection and localization method based on multi-granularity representation decoupling and collaboration

Yuan Chengsheng[1,2], Cao Fugui[1,2], Wang Yili[1,2], Cao Yi[3], Liu Qingcheng[1], Fu Zhangjie[1,2]
1. School of Cyber Science and Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China
2. Ministry of Education Engineering Research Center for Digital Forensics, Nanjing University of Information Science and Technology, Nanjing 210044, China
3. School of Internet of Things Engineering, Wuxi University, Wuxi 214105, China

Abstract: Images and other multimedia content transmitted over mobile broadband are vulnerable to interception and tampering, which may compromise the authenticity of communication content. To address insufficient cross-dataset generalization and inaccurate localization of complex boundaries and small tampered regions, a highly generalizable image tampering detection and localization method based on multi-granularity representation decoupling and collaboration was proposed. Firstly, fine-grained artifact and coarse-grained structural representations were constructed to capture local forensic cues and regional structural information, respectively. Secondly, an explicit decoupling constraint was introduced to reduce redundant correlations between representations of different granularities. Finally, an artifact-prior-guided gated adaptive fusion module dynamically balanced local details and contextual structural information. Experiments on CASIA_V1, Columbia, NIST16, and Coverage show that the proposed method achieves competitive overall performance in cross-dataset testing, with strong generalization ability and localization accuracy.

Key words: mobile broadband, image tampering detection and localization, multi-granularity representation, explicit decoupling, gated adaptive fusion

收稿日期：2026–04–21；**修回日期：**2026–06–17
通信作者：曹燚，caoyi@cwxu.edu.cn
基金项目：国家自然科学基金资助项目（No.U22B2062, No.U23B2023）；江苏省研究生创新计划项目（No.SUCX25_0522）
Foundation Items: The National Natural Science Foundation of China (No.U22B2062, No.U23B2023), The Postgraduate Research and Practice Innovation Program of Jiangsu Province (No.SUCX25_0522)

0　引言

随着5G通信技术的发展与移动互联网应用的普及，数字图像已成为网络空间中传播最为广泛的多媒体信息形式之一，在新闻传媒、电子商务、公共安全及司法取证等领域发挥着重要作用。然而，简易图像编辑工具的广泛使用显著降低了图像伪造的技术壁垒，使得篡改操作更为便捷且隐蔽。一旦恶意篡改的图像通过社交网络等渠道传播，不仅会误导公众舆论、损害个人隐私与企业利益，更可能对社会信任体系与国家安全构成潜在威胁。因此，研究能够准确检测并定位篡改区域的图像取证方法，对于增强多媒体内容真实性验证能力具有重要意义。

当前常见的图像篡改方式主要包括拼接、复制粘贴和删除修复等。这类操作虽然在视觉表现上各不相同，但通常会破坏图像在成像、压缩和编辑过程中形成的统计一致性，并留下重采样痕迹、噪声异常、光照不一致或局部结构矛盾等可供取证的线索。早期研究主要依赖手工设计的特征来实现篡改检测与定位。例如，Farid[1]从物理光照与几何一致性角度构建了多维取证特征空间；Popescu等[2]利用线性插值引入的周期性特征检测重采样痕迹，实现对缩放与拼接篡改的定位；Fridrich等[3]通过基于块匹配的算法为复制粘贴篡改检测奠定技术基础。这类方法具有较强的可解释性，但通常依赖特定篡改类型或人工统计假设，在面对复杂背景、多源图像以及后处理干扰等条件时，其泛化能力有限。

随着深度学习的发展，图像篡改检测与定位逐渐转化为像素级密集预测任务。现有方法主要从特征建模、边界约束、多尺度融合与跨域泛化等方面提升定位性能。例如，王珠珠[4]基于U型检测网络提升了篡改区域的检测与定位能力。Bappy等[5]利用卷积神经网络–长短期记忆网络（convolutional neural network-long short-term memory network，CNN-LSTM）结构对边界邻域差异进行建模。随后，Zhou等[6]提出双流特征学习框架，联合利用红绿蓝（red-green-blue，RGB）颜色空间内容与噪声线索增强对篡改区域的检测与定位能力。Wu等[7]提出篡改追踪网络（manipulation tracing network，ManTra-Net），将篡改定位视为异常检测问题，引入局部异常评估机制，使模型能够对不同类型伪造痕迹形成较为一致的响应。在此基础上，Zhou等[8]提出生成–分割–细化网络（generate-segment-refine network，GSR-Net），采用生成、分割、细化的多阶段结构提升通用篡改分割性能。为缓解跨数据集测试中模型易受语义内容干扰的问题，Chen等[9]提出多视角多尺度监督网络（multi-view multi-scale supervised network，MVSS-Net），结合多视角特征学习与多尺度监督增强边界伪迹和噪声异常感知。针对更加贴近实际应用的真实场景，朱叶等[10]提出高分辨率表征扩张注意力网络（high-resolution representation dilation attention network，HRDA-Net），面向真实场景下的多篡改检测与定位任务进行建模。此后，Zhou等[11]进一步提出非互斥对比学习（non-mutually exclusive contrastive learning，NCL），通过NCL改善边界区域训练稳定性。此外，Kwon等[12]提出压缩伪迹追踪网络（compression artifact tracing network，CAT-Net），利用RGB颜色空间与离散余弦变换（DCT）双域建模压缩伪迹以提升拼接定位性能。Wang等[13]提出对象级Transformer网络（ObjectFormer），从对象级一致性关系出发增强篡改区域建模。Liu等[14]提出渐进式空间–通道相关网络（progressive spatio-channel correlation network，PSCC-Net），通过渐进式空间通道相关性学习进一步改善细粒度定位能力。上述方法有效推动了图像篡改定位性能的提升，但在复杂跨域场景下仍存在进一步的改进空间。

然而，现有方法在复杂场景下仍面临两个关键问题。一方面，模型的泛化能力不足。不同数据集在成像设备、场景内容、编辑方式和后处理操作上存在明显差异，模型易学习到与数据分布或场景内容相关的偏置信息，导致跨数据集、跨篡改类型测试中出现误检或漏检。已有研究表明，压缩、模糊等后处理会进一步削弱篡改痕迹，增加了检测与定位的难度，谭舜泉等[15]的研究也证实此类场景会显著增加稳定定位的难度。另一方面，模型的精细定位能力仍有待提升。复杂边界与小尺寸篡改对模型的精细定位能力提出了更高要求。尽管扩大感受野和引入多尺度特征有助于增强整体区域感知，但在逐级下采样和跨尺度融合过程中，边界邻域的局部异常、高频统计突变和小区域篡改线索仍容易被削弱，从而造成边界模糊、区域不连续或小区域漏检。

从任务本质上看，图像篡改检测与定位并非一般意义上的分割问题，其重点并非识别图像中的目标类别，而在于发现篡改操作对图像一致性关系造成的破坏。这种破坏既可能表现为边界过渡、纹理断裂、噪声失衡和压缩异常等细粒度伪迹，也可能表现为篡改区域内部结构连续性、区域一致性和上下文关系的异常。因此，模型既需要捕获局部伪迹线索，也需要建模区域结构线索。若过度依赖局部细节，模型易受复杂背景纹理或弱伪迹样本的干扰；若过度依赖高层上下文，又可能忽视边界邻域与小区域篡改中的细微异常。进一步来看，在统一像素级监督下，不同粒度特征往往会共同响应显著内容区域，导致细粒度伪迹表征与粗粒度结构表征之间出现功能重叠和信息冗余，从而削弱二者的互补作用。因此，如何在局部伪迹感知与区域一致性建模之间建立有效的分工与协同关系，是提升模型跨域泛化能力和精细定位能力的关键。

针对上述问题，本文提出一种基于多粒度表征解耦与协同的高泛化图像篡改检测与定位方法。该方法首先分别构建细粒度异常表征与粗粒度结构表征，以刻画局部伪迹线索和区域一致性信息。随后引入显式解耦约束，在归一化嵌入空间中降低不同粒度表征之间的冗余相关性，促使两类表征形成更加清晰的功能分工。最后设计伪迹先验引导的门控自适应融合模块，根据不同空间位置对局部细节和上下文结构的需求差异，动态调节多粒度特征的融合权重。本文在多个公开数据集上进行跨数据集测试，并通过消融实验和鲁棒性实验验证所提方法的有效性。

1　模型架构

1.1　模型整体框架

本文模型的整体架构如图 1 所示。基于前文对任务本质的分析，图像篡改定位是对篡改操作引起的局部伪迹异常与区域一致性破坏进行联合识别。为此，本文将任务建模为细粒度伪迹感知与粗粒度结构建模的分工协同问题，并在整体框架中构建多粒度表征、显式解耦和门控自适应融合 3 个阶段。

首先，给定输入图像，模型利用稀疏视觉 Transformer（vision transformer，ViT）编码器提取多层级特征，并分别构建细粒度伪迹表征与粗粒度结构表征。其中，细粒度伪迹表征主要面向边界过渡、纹理断裂、噪声异常和小区域篡改等局部线索，用于增强模型对弱伪迹和复杂边界的感知能力；粗粒度结构表征则侧重刻画篡改区域内部一致性、结构连续性和上下文关系，用于提升预测区域的完整性与稳定性。通过这种多粒度建模方式，模型能够同时关注局部异常响应和整体区域结构，避免仅依赖单一尺度特征导致的边界模糊或区域响应不连续问题。

随后，考虑到两类表征在统一像素级监督下可能共同响应显著内容区域，进而产生功能重叠和信息冗余，本文引入显式解耦模块对细粒度伪迹表征与粗粒度结构表征施加嵌入空间约束。该约束并非

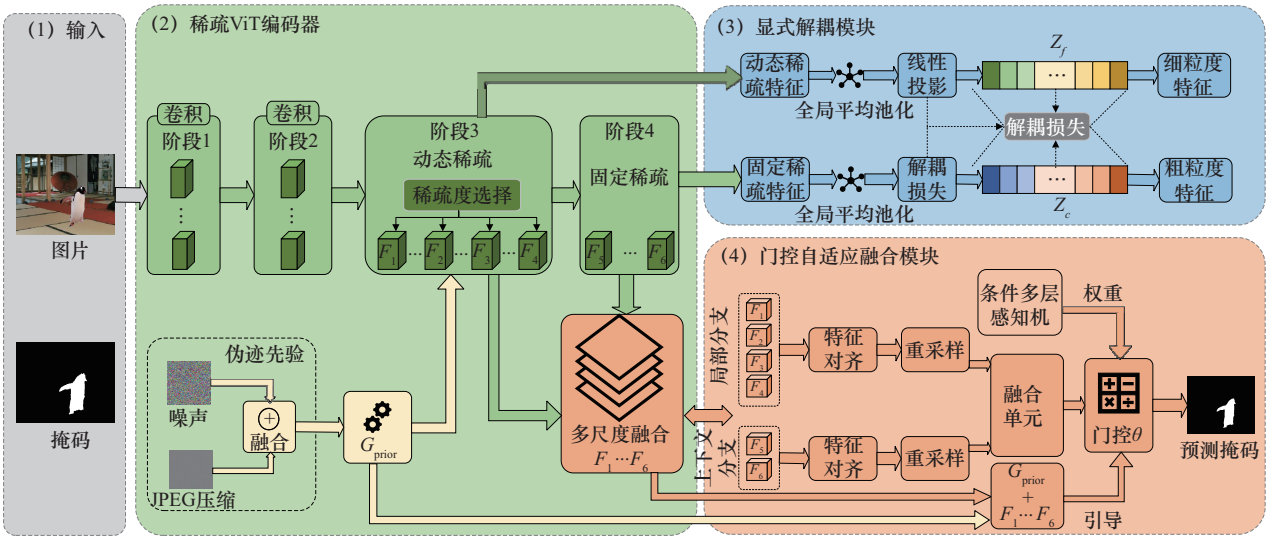


图1　本文模型的整体架构

简单增加网络分支，而是通过降低不同粒度表征之间的冗余相关性，促进二者在局部伪迹刻画和区域结构建模中的功能分工，为后续协同融合提供互补性更强的特征输入。

最后，针对不同空间位置对局部细节和上下文结构需求不同的问题，本文设计伪迹先验引导的门控自适应融合模块，对解耦后的多粒度表征进行协同融合。具体而言，在边界复杂、伪迹响应明显或小区域篡改位置，模型倾向于保留更多细粒度伪迹信息。在区域内部或结构连续性较强的位置，模型则更多利用粗粒度结构信息，以保持预测结果的整体稳定性。相较于直接拼接、逐元素相加或统一注意力融合等常规多尺度融合方式，所提方法更强调局部伪迹线索与区域结构线索之间的功能分工和空间自适应协同，从而为兼顾边界刻画精细性与区域响应稳定性提供更为合理的建模依据。

1.2　稀疏 ViT 编码器

图像篡改检测与定位既依赖边界邻域中的纹理断裂、噪声失衡、压缩不一致等局部伪迹，也依赖篡改区域内部结构连续性和上下文一致性，仅依赖单一粒度特征难以同时兼顾边界敏感性与区域稳定性。为此，本文构建了稀疏 ViT 编码器[16]，并在编码过程中引入伪迹先验引导的稀疏注意机制，使网络能够依据输入内容及伪迹响应自适应调节建模范围。具体而言，编码器前两阶段主要完成基础视觉特征提取，为后续伪迹响应建模和上下文聚合提供底层表示。第三阶段采用动态稀疏注意机制，根据局部伪迹响应自适应选择稀疏度，以增强模型对边界邻域、纹理突变和局部异常线索的感知能力。第四阶段采用固定稀疏注意机制，聚合区域级上下文信息，提升对篡改区域内部一致性和整体结构关系的表征能力。通过上述设计，编码过程逐步生成具有不同功能侧重的多粒度特征表示，为后续表征解耦与协同融合提供基础。其中，用于构建伪迹先验的浅层特征记为 $\boldsymbol{F}_{\mathrm{p}}$。考虑到噪声扰动[17]与压缩失真[18]是两类具有代表性且互为补充的篡改痕迹，本文分别构建噪声响应分支和压缩响应分支，并将二者融合为统一的伪迹先验 $\boldsymbol{G}$。

$$\boldsymbol{G}_{\mathrm{n}} = \psi_{\mathrm{n}}\left(\boldsymbol{F}_{\mathrm{p}}\right) \tag{1}$$

$$\boldsymbol{G}_{\mathrm{j}} = u\left(\left[\, p\left(\psi_{\mathrm{j}}\left(\boldsymbol{F}_{\mathrm{p}}\right)\right)\right]\right) \tag{2}$$

$$\boldsymbol{G} = \psi_{\mathrm{f}}\left(\left[\boldsymbol{G}_{\mathrm{n}}, \boldsymbol{G}_{\mathrm{j}}\right]\right) \tag{3}$$

其中，$\boldsymbol{G}_{\mathrm{n}}$ 和 $\boldsymbol{G}_{\mathrm{j}}$ 分别表示噪声响应与压缩响应，$\psi_{\mathrm{n}}(\cdot)$ 和 $\psi_{\mathrm{j}}(\cdot)$ 为对应的映射函数，$p(\cdot)$ 与 $u(\cdot)$ 分别表示池化与上采样操作，$[\cdot]$ 表示通道拼接。此外，$\boldsymbol{G}$ 不直接给出篡改位置的确定性判断，而是为后续粒度选择提供与输入内容相关的伪迹提示。之后，对第 t 个自适应稀疏编码单元的输入特征 $\boldsymbol{X}_t$，其中稀疏选择器的选择依据输入内容与伪迹先验联合估计稀疏级别，并据此完成注意力建模，可表示为：

$$\boldsymbol{p}_t = S_t(\boldsymbol{X}_t, \boldsymbol{G}) \sum_{m=1}^{M} p_{t,m} = 1 \tag{4}$$

$$r_t = Q\left(\sum_{m=1}^{M} p_{t,m} r_m\right) \boldsymbol{X}_{t+1} = A\left(\boldsymbol{X}_t;\, r_t\right) \tag{5}$$

其中，$S_t(\cdot)$ 表示基于候选稀疏级别集合的自适应选择函数，$p_t=[p_{t,1},\cdots,p_{t,m}]$ 表示各候选稀疏级别对应的选择概率，$Q(\cdot)$ 表示量化操作，$A(\cdot)$ 表示由稀疏级别 r_t 控制的注意力变换。该设计使模型能够根据输入内容和伪迹先验自适应调整建模粒度，从而在不同图像内容、伪迹强度及结构复杂度下选择相应的局部建模范围。为进一步抑制相邻编码单元之间稀疏度的剧烈波动，对连续编码单元的稀疏级别施加单调约束，以增强粒度选择的稳定性。

经上述过程得到的中层自适应稀疏表征记为 $\boldsymbol{F}_{\mathrm{f}}$，其主要关注边界过渡、纹理断裂及统计异常等细粒度线索。由深层编码获得的上下文表征记为 $\boldsymbol{F}_{\mathrm{c}}$，其侧重于区域一致性、结构约束及全局语义信息。由此，网络在编码阶段形成了具有明确分工的多粒度表征[19]，也为后续显式解耦与门控自适应融合提供了基础。

1.3　显式解耦模块

尽管多粒度表征能够同时引入局部伪迹信息和区域上下文信息，但仅构建多分支特征并不意味着二者一定能够形成有效互补。在统一像素级定位监督下，细粒度表征与粗粒度表征可能共同响应图像中的显著内容区域，从而产生表征重叠、信息冗余和功能分工不清的问题。具体而言，粗粒度分支在学习区域一致性时可能混入过多局部纹理和边界细节，而细粒度分支也可能受到区域结构响应的影响，导致其对局部伪迹和边界异常的刻画能力下

降。这样一来，多粒度建模容易退化为多分支特征堆叠，难以充分发挥局部伪迹线索与区域结构线索之间的互补作用。针对这一问题，本文设计显式解耦模块，在嵌入空间中对两类表征施加去冗余约束，以减少重复编码，并促进细粒度表征与粗粒度表征形成更加清晰的功能侧重。

将细粒度表征和粗粒度表征分别记为$\boldsymbol{F}_{\mathrm{f}}$和$\boldsymbol{F}_{\mathrm{c}}$。首先，通过全局平均池化操作以及相互独立的投影头，将两类特征映射到统一嵌入空间，并对其进行归一化处理。

$$\boldsymbol{z}_{\mathrm{f}} = \mathrm{Norm}_2\left(\varphi_{\mathrm{f}}\left(\mathrm{GAP}\left(\boldsymbol{F}_{\mathrm{f}}\right)\right)\right) \tag{6}$$

$$\boldsymbol{z}_{\mathrm{c}} = \mathrm{Norm}_2\left(\varphi_{\mathrm{c}}\left(\mathrm{GAP}\left(\boldsymbol{F}_{\mathrm{c}}\right)\right)\right) \tag{7}$$

其中，GAP(·)表示全局平均池化，$\varphi_{\mathrm{f}}(\cdot)$和$\varphi_{\mathrm{c}}(\cdot)$分别表示细粒度分支与粗粒度分支的投影映射，$\mathrm{Norm}_2(\cdot)$表示$L_2$归一化操作。经过该处理后，$\boldsymbol{z}_{\mathrm{f}}$更集中地表征与边界、伪迹和局部异常相关的信息，$\boldsymbol{z}_{\mathrm{c}}$更强调区域一致性、上下文依赖与全局结构约束。为了进一步降低两类表征之间的冗余耦合，在归一化嵌入空间中最小化二者的相似性，显式解耦损失定义为：

$$L_{\mathrm{dec}} = \frac{1}{B}\sum_{b=1}^{B}\left(\left(\boldsymbol{z}_{\mathrm{f}}^{(b)}\right)^{\mathrm{T}}\boldsymbol{z}_{\mathrm{c}}^{(b)}\right)^2 \tag{8}$$

其中，B表示批量大小，$\boldsymbol{z}_{\mathrm{f}}^{(b)}$和$\boldsymbol{z}_{\mathrm{c}}^{(b)}$分别表示第$b$个样本对应的细粒度嵌入和粗粒度嵌入。由于$\boldsymbol{z}_{\mathrm{f}}$和$\boldsymbol{z}_{\mathrm{c}}$已进行$L_2$归一化，使得约束二者的余弦相似度尽可能接近于0，从而促使细粒度分支与粗粒度分支学习到更具差异性的表示，减少无效的重复编码。

同时，本文的显式解耦模块并非严格意义上的互信息最小化，而是从抑制多粒度表征冗余的角度出发，构建一种基于归一化嵌入空间的去相关约束机制。余弦相似度刻画了特征向量之间的方向相关性，属于内积相似性度量，主要用于降低细粒度表征与粗粒度表征之间的线性相关冗余。与互信息神经估计等[20]方法不相同，该约束并不直接估计两类随机变量之间的互信息，也不刻画二者之间全部的统计依赖关系，尤其是复杂的非线性依赖关系，而是侧重于抑制多粒度表征之间的重复编码和耦合响应，并促进细粒度伪迹表征与粗粒度结构表征形成差异化表达，从而在保留篡改定位判别信息的基础上增强二者的互补性。

与近期基于充分性与最小性约束的信息论图像伪造定位框架（SUMI-IFL）从信息充分性与最小性角度对篡改相关信息进行建模不同，本文关注的是多粒度定位线索之间的功能分化与协同利用。SUMI-IFL主要通过信息充分性约束保留篡改判别所需信息，并通过最小性约束抑制任务无关信息，从而获得紧凑的篡改相关表征。本文的显式解耦模块主要作用于细粒度伪迹表征与粗粒度结构表征之间，通过嵌入空间中的去相关约束降低二者之间的冗余响应。与SUMI-IFL不同，本文的细粒度表征与粗粒度表征虽具有不同的建模侧重点，但本质上均为同一篡改事件在不同粒度上的表征形式，并最终共同服务于像素级定位目标。与追求二者完全独立不同，本文在保留必要共享判别信息的前提下减少功能重叠，使细粒度分支更关注边界过渡、纹理断裂、噪声异常及小区域篡改线索，而粗粒度分支更关注区域一致性与结构连续性，从而为后续门控自适应融合提供互补性更强的特征输入。

1.4 门控自适应融合模块

在获得具有明确分工的多粒度表征后，解码阶段的协同方式直接影响模型预测结果的边界清晰度与区域完整性。常见的静态融合方式，如简单拼接、逐元素相加或固定比例加权，难以满足不同空间位置对局部细节和全局上下文的差异化需求，容易导致局部细节被稀释或区域响应不稳定。为此，本文还设计了门控自适应融合模块，通过伪迹先验引导与空间相关的门控机制，实现多粒度特征的动态协同融合。

首先，将不同层级的多路特征映射到统一的通道数与空间尺度，记对齐后的特征集合为$\{\boldsymbol{F}_i\}$。在此基础上，结合各路特征的全局统计信息和伪迹先验$\boldsymbol{G}$生成动态权重，并分别构造局部分支表示与上下文分支表示：

$$\boldsymbol{w} = W\left(\left\{\boldsymbol{F}_i\right\}_{i=1}^{N}, \boldsymbol{G}\right) \tag{9}$$

$$\boldsymbol{F}_{\mathrm{loc}} = \sum_{i \in S_{\mathrm{loc}}} w_i * \boldsymbol{F}_i \tag{10}$$

$$\boldsymbol{F}_{\mathrm{ctx}} = \sum_{i \in S_{\mathrm{ctx}}} w_i * \boldsymbol{F}_i \tag{11}$$

$$\boldsymbol{F}_{\mathrm{b}} = \boldsymbol{F}_{\mathrm{loc}} + \boldsymbol{F}_{\mathrm{ctx}} \tag{12}$$

其中，$W(\cdot)$表示动态权重预测函数，S_{loc}和S_{ctx}分别表示低层细节特征索引集合与高层上下文特征索引集合，*表示逐元素乘法。由此，$\boldsymbol{F}_{\mathrm{loc}}$更侧重局部纹理、边界过渡和细粒度伪迹信息，$\boldsymbol{F}_{\mathrm{ctx}}$更侧重区域一致性和全局语义约束，$\boldsymbol{F}_{\mathrm{b}}$为二者的基础融合结果。之后利用伪迹先验和基础融合结果构造门控引导图。

$$\boldsymbol{P} = \mathrm{Sigmoid}\left(C_{\mathrm{p}}\left(\left[\mathrm{Up}(\boldsymbol{G}),\boldsymbol{F}_{\mathrm{b}}\right]\right)\right) \tag{13}$$

其中，$C_{\mathrm{p}}(\cdot)$表示映射函数，$\mathrm{Up}(\cdot)$表示上采样操作，用于将伪迹先验$\boldsymbol{G}$调整至与基础融合特征$\boldsymbol{F}_{\mathrm{b}}$一致的空间尺寸，Sigmoid(·)表示Sigmoid激活函数。$\boldsymbol{P}$反映了不同空间位置对局部细节信息和上下文信息的潜在需求差异，从而为后续门控分配提供有效引导。在此基础上，门控自适应融合模块分别对局部分支与上下文分支进行映射，并生成空间相关的门控权重。

$$\boldsymbol{H}_{\mathrm{loc}} = C_{\mathrm{loc}}\left(\boldsymbol{F}_{\mathrm{loc}}\right) \quad \boldsymbol{H}_{\mathrm{ctx}} = C_{\mathrm{ctx}}\left(\boldsymbol{F}_{\mathrm{ctx}}\right) \tag{14}$$

$$\boldsymbol{M} = \mathrm{Sigmoid}\left(\mathrm{Gate}\left(\left[\boldsymbol{H}_{\mathrm{loc}},\boldsymbol{H}_{\mathrm{ctx}},\boldsymbol{P}\right]\right)\right) \tag{15}$$

$$\boldsymbol{F} = C_{\mathrm{o}}\left(\boldsymbol{F}_{\mathrm{b}} + \boldsymbol{M}*\boldsymbol{H}_{\mathrm{loc}} + (1-\boldsymbol{M})*\boldsymbol{H}_{\mathrm{ctx}}\right) \tag{16}$$

其中，$C_{\mathrm{loc}}(\cdot)$、$C_{\mathrm{ctx}}(\cdot)$和$C_{\mathrm{o}}(\cdot)$分别表示局部分支映射、上下文分支映射和输出映射函数，Gate(·)表示门控权重预测函数，$\boldsymbol{M}$为门控权重图。在靠近篡改边界、局部结构复杂或伪迹响应显著的位置，门控提升局部分支的权重。在更依赖整体结构约束与区域稳定性的位置，模型保留更多上下文分支的信息。与固定融合方式相比，该模块在空间维度上实现了更灵活的差异化特征分配[21]。

为避免训练初期门控权重过早偏向某一分支，本文进一步引入门控正则项。

$$L_{\mathrm{gate}} = E\left[(\boldsymbol{M}-0.5)^2\right] \tag{17}$$

该正则项用于约束门控权重，避免其过早偏离中性状态，以提升融合过程的稳定性，减轻模型对单一路径的过度依赖，使局部伪迹信息与上下文信息能够逐步形成稳健的协同关系。综合主定位目标、显式解耦约束与门控正则项，检测模型的总体损失记为：

$$L = L_{\mathrm{loc}} + a_{\mathrm{dec}}L_{\mathrm{dec}} + a_{\mathrm{gate}}L_{\mathrm{gate}} \tag{18}$$

其中，L_{loc}表示主定位损失，a_{dec}和a_{gate}为平衡系数。

值得注意的是，本文采用伪迹先验引导的门控自适应融合，并不是简单增加一个融合结构，而是针对多粒度特征在解码阶段如何分配、如何协同的问题进行建模。对于图像篡改检测定位任务，细粒度特征更容易捕获边界断裂、纹理突变、噪声异常和压缩不一致等局部伪迹，但也更容易受到自然纹理和背景噪声干扰。粗粒度特征能够提供区域一致性和上下文结构约束，有助于保持预测区域完整性，但在聚合过程中可能削弱边界细节和小区域篡改响应。因此，两类特征并不存在固定的主次关系，关键在于根据不同空间位置的伪迹强度和结构需求，自适应决定局部细节与上下文信息的贡献比例。本文引入的伪迹先验并不直接作为篡改掩码，而是作为一种任务相关的软引导，使门控权重在学习过程中能够参考噪声失衡、压缩异常和纹理不连续等取证线索，从而使融合过程更贴近篡改定位任务本身。

相比之下，逐元素相加默认不同粒度特征在所有位置具有相近作用，容易造成局部伪迹被高层上下文平滑，或低层噪声干扰区域预测。通道拼接虽然保留了更多信息，但本质上主要是维度扩展，后续解码器仍需从冗余特征中重新筛选有效线索，增加了学习负担，也可能引入无关背景响应。无自适应融合虽然具备一定自适应能力，但门控权重完全由特征自身学习，在跨数据集或复杂背景下容易受到场景语义、目标外观和数据分布偏置影响，未必能稳定聚焦真实篡改痕迹。本文的自适应融合则在空间位置上对局部分支和上下文分支进行动态分配，既能减少相加和拼接带来的无选择融合问题，也能降低无自适应融合对纯数据驱动学习的依赖，从而在边界刻画、区域完整性和跨域稳定性之间取得更好的平衡。

2　实验结果

2.1　数据集介绍

本文采用多源数据联合训练、跨数据集测试的实验设置。训练集由TampCOCO、CASIA_V2、IMD2020和FantasticReality_v1组成，测试集包括Coverage、Columbia、CASIA_V1和NIST16。采用这种设计，一方面多源训练数据能够覆盖拼接、复制粘贴、删除以及边界模糊、颜色不一致等多种篡改形式，有助于模型学习到具有共性的篡改特征。

另一方面，在与训练集分布差异明显的公开测试集上进行评估，能够更客观地反映模型的跨数据集泛化能力。若模型过度依赖特定数据集中的纹理统计或语义线索，则难以在分布差异显著的测试集上保持稳定输出，因此该实验设置更适合验证方法的实际应用价值与泛化水平。

训练集方面，TampCOCO[22]是一个大规模合成篡改数据集，由多个子集构成。其中，sp子集主要对应拼接篡改，包含199 999张篡改图像；cm子集主要对应复制粘贴篡改，包含199 429张篡改图像；bcm子集主要对应边界模糊的复制粘贴篡改，包含199 443张篡改图像；bcmc子集主要对应边界模糊且颜色不一致的复制粘贴篡改，包含199 443张篡改图像。上述4个子集在边缘过渡方式和局部统计特征上存在较大差异，能够为模型提供大规模、多样化的监督样本。CASIA_V2[23]是图像篡改检测与定位研究中广泛采用的公开数据集，包含5 123张篡改图像。该数据集主要分为拼接和复制粘贴两类篡改，同时部分样本还经过裁剪、旋转、模糊等后处理，具有较强的代表性。IMD2020[24]包含2 010张篡改图像，篡改类型主要包括拼接、复制粘贴和删除，且部分样本来源于真实网络图像，场景更接近实际应用。FantasticReality_v1[25]包含16 592张真实图像和19 423张篡改图像，主要面向拼接与合成篡改场景构建，能够进一步补充更自然的前景融合、光照变化与语义内容差异，有助于提升模型对复杂场景的适应能力。

测试集方面，Coverage[26]主要用于评估复制粘贴篡改的检测与定位性能，本文使用其中的100张篡改图像。Columbia[27]是经典的拼接篡改测试集，本文使用其中180张篡改图像。CASIA_V1共包含919张篡改图像，篡改类型主要包括拼接和复制粘贴，同时部分图像伴随裁剪、旋转、模糊等后处理操作。NIST16[28]共包含564张篡改图像，涵盖拼接、复制粘贴和删除等多种篡改方式，且图像来源与编辑条件更为复杂。将上述4个数据集共同作为测试集，可以从复制粘贴、拼接以及复杂真实编辑等多个角度，对模型的定位能力与跨数据集泛化性能进行较为全面的评估。

配置环境方面，本文方法基于PyTorch框架[29]实现，在Tesla P40上进行训练与测试。训练批大小为6，测试批大小为4，采用梯度累积策略，累积步数为16。模型共训练68个训练轮次（epoch），每4个epoch进行一次测试评估。优化器采用AdamW[30]，权重衰减系数设为0.05，初始学习率设为1×10^{-4}，最小学习率设为1×10^{-6}，预热轮数设为5，温度参数τ从2.0逐步退火至0.4。

2.2 评价指标

本文采用像素级F1分数（F1-score）、交并比（intersection over union, IoU）以及接收者操作特征（receiver operating characteristic, ROC）曲线下面积（area under the curve, AUC）作为模型性能评价指标。上述指标分别从像素分类准确性、预测区域与真实区域的重叠程度以及不同判别阈值下的整体区分能力，对模型的检测与定位结果进行综合评估。实验中篡改区域为正类，真实区域为负类。在像素级二分类任务中，TP、FP、TN和FN分别表示被正确判定为篡改的像素数、被错误判定为篡改的像素数、被正确判定为真实的像素数以及被错误判定为真实的像素数。精确率（Precision）与召回率（Recall）定义为：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{19}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{20}$$

其中，精确率表示模型预测为篡改的像素中真实篡改像素所占的比例，召回率表示真实篡改像素中被模型正确检出的比例。在此基础上，F1分数定义为精确率与召回率的调和均值，即

$$\text{F1-score}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{21}$$

F1分数能够同时兼顾查准率与查全率，因此适合用于评价篡改检测与定位任务中模型对篡改区域的综合检出能力。只有当模型同时具有较高的精确率和召回率时，F1分数才会相应较高。IoU用于衡量预测篡改区域与真实篡改区域之间的重叠程度，其定义为：

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{22}$$

与F1分数相比，IoU对预测区域与真实区域的重叠质量要求更高，更适合反映模型在区域级定位上的准确程度。当预测掩码与真实掩码越接近时，IoU取值越大。除固定阈值下的评价指标外，本文

还采用 AUC 作为补充指标。AUC 来源于 ROC 曲线，该曲线以假正例率（false positive rate, FPR）为横轴，真正例率（true positive rate, TPR）为纵轴，AUC 则表示 ROC 曲线下的面积，可表示为：

$$\mathrm{AUC}=\int_0^1 \mathrm{TPR}(\mathrm{FPR})\,\mathrm{d}(\mathrm{FPR}) \tag{23}$$

AUC 不依赖单一固定阈值，而是统计模型在不同判别阈值下的整体分类能力，能够更全面地反映模型输出概率图的可分性与稳定性。AUC 越大，说明模型对篡改像素与真实像素的区分能力越强。本文实验 F1 分数和 IoU 均在固定阈值 0.5 下进行统计，即当像素预测概率大于等于 0.5 时，将其判定为篡改像素，否则判定为真实像素。综合来看，F1 分数更能体现模型在精确率和召回率之间的平衡，IoU 更关注预测区域与真实区域之间的重叠质量，AUC 则能够从多阈值角度补充评价模型的整体判别性能。因此，本文同时采用这 3 个指标，以对所提方法的篡改检测与定位能力进行更加全面和客观的分析。

2.3　跨数据集测试与泛化性分析

本文在上文介绍的 4 个公开测试数据集上进行了测试，并与可信图像伪造检测与定位框架（TruFor）、决策环境构建框架（construct decision-making environment，CoDE）、高效图像篡改定位网络（effective image tampering localization network，EITLNeT）、稀疏视觉 Transformer（sparse vision transformer，SparseViT）、多样化数据生成取证框架（DDG）、介观多尺度混合架构（Mesorch）、噪声感知对比定位网络（noise-aware contrastive localization network，NCNeT）、聚焦区域差异网络（focus region discrepancy network，FRD-Net）、协同 Transformer（collaborative transformers，Co-Transformers）、边缘感知重建引导网络（edge-aware reconstruction-guided network，EARG-Net）等代表性方法在同一实验条件下进行对比。评价指标包括固定阈值 0.5 下的像素级 F1 分数和 IoU，同时引入 AUC 作为补充指标。为便于比较，表 1 和表 2 中各列最优结果以粗体表示，次优结果以下划线表示。对于部分近期方法未公开或未能在统一实验设置下获得的指标结果，表中以“—”标记，在指标比较时不参与判断。

由表 1 可以看出，本文方法在跨数据集测试中整体表现较为稳定，平均 F1 分数达到 0.714，平均 IoU 达到 0.663，综合指标优于多数对比方法。尤其在 CASIA_V1 数据集上，本文方法取得了 0.838 的 F1 分数与 0.780 的 IoU，AUC 进一步达到 0.988，表明在拼接与复制粘贴并存且存在一定后处理干扰的经典测试集上，本文方法能够较好地保持像素级判别能力与区域定位稳定性。与 SparseViT、Mesorch 和 NCNeT 等方法相比，本文方法并非单独强化非语义特征、混合尺度结构或噪声交互，而是通过多粒度表征解耦降低局部伪迹线索与区域结构线索之

表 1　不同模型在测试集上固定阈值 0.5 时的 F1 分数与 IoU 比较

方法	CASIA_V1		Columbia		NIST16		Coverage		平均	
	F1 分数	IoU	F1 分数	IoU	F1 分数	IoU	F1 分数	IoU	F1 分数	IoU
TruFor[31]	0.818	0.764	0.885	0.859	0.348	0.301	0.457	0.419	0.627	0.586
CoDE[32]	0.723	0.637	0.881	0.844	0.420	0.339	0.464	0.362	0.622	0.546
EITLNeT[33]	0.530	0.492	0.881	0.851	0.308	0.256	0.448	0.371	0.542	0.493
SparseViT[34]	0.827	0.775	0.959	0.938	0.384	0.331	0.513	0.472	0.671	0.629
DDG[35]	0.742	0.675	0.910	0.877	0.377	0.302	0.438	0.367	0.617	0.555
Mesorch[36]	**0.839**	**0.787**	0.890	0.876	0.392	0.342	0.585	0.536	0.676	0.635
NCNeT[37]	0.823	0.761	0.932	0.903	0.402	0.327	0.584	0.497	0.682	0.622
FRD-Net[38]	0.829	—	**0.973**	—	0.412	—	0.601	—	0.703	—
Co-Transformers[39]	0.807	—	0.941	—	0.425	—	0.634	—	0.701	—
EARG-Net[40]	0.498	—	0.924	—	**0.589**	—	**0.755**	—	0.691	—
本文方法	0.838	0.780	0.953	**0.940**	0.461	**0.389**	0.603	**0.543**	**0.714**	**0.663**

间的冗余耦合，因此在CASIA_V1这类篡改类型较为混合的数据集上表现出较好的均衡性。

表2　不同模型在测试集上的AUC比较

方法	CASIA_V1	Columbia	NIST16	Coverage	平均
TruFor	0.897	0.899	0.845	0.846	0.872
CoDE	0.921	0.934	0.839	0.872	0.892
EITLNeT	0.873	0.918	0.826	0.839	0.864
SparseViT	0.982	0.970	0.861	0.935	0.937
DDG	0.949	0.947	0.819	0.858	0.893
Mesorch	0.985	0.909	0.888	0.932	0.929
NCNeT	0.977	0.951	0.872	0.947	0.936
FRD-Net	—	—	—	—	—
Co-Transformers	—	—	—	—	—
EARG-Net	0.840	**0.995**	0.881	**0.988**	0.920
本文方法	**0.988**	0.957	**0.889**	0.964	**0.949**

在Columbia数据集上，FRD-Net和SparseViT取得了较高的F1分数，这与该数据集以拼接篡改为主、目标边界和区域差异相对清晰有关。FRD-Net通过放大宏观语义差异与微观结构差异，增强真实区域与篡改区域之间的不一致性，SparseViT则利用稀疏注意力机制削弱语义依赖，突出非语义层面的篡改痕迹，因此在拼接边界较明确的场景中易于获得较高的固定阈值响应。本文方法的F1分数为0.953，略低于上述方法，但IoU达到0.940，在所有对比方法中处于最优水平，表明本文方法预测的篡改区域与真实掩码具有更好的空间重叠质量，区域范围控制更为稳定。

在NIST16数据集上，本文方法的F1分数低于EARG-Net，但高于FRD-Net、Co-Transformers、SparseViT等方法，同时取得了0.389的IoU和0.889的AUC。NIST16数据集包含拼接、复制粘贴、删除修复及复杂后处理等情况，篡改痕迹常表现为弱边界、融合边缘或局部修复不一致。EARG-Net利用预训练图像修复模型的自然图像先验，通过边缘感知重建和原图–重建图残差来放大异常区域，因此在此类修复型、弱边界样本上更易获得较高的召回率，从而提升固定阈值下的F1分数。相比之下，本文方法更强调局部伪迹与区域结构的协同建模，输出响应相对平滑和保守，因此在固定阈值0.5以下的F1分数不占优势。然而，本文方法的AUC高于EARG-Net，表明本文方法对篡改像素与真实像素仍具有较好的整体区分能力，主要差异体现在固定阈值下的响应强度，而非判别能力不足。

在Coverage数据集上，EARG-Net和Co-Transformers的F1分数高于本文方法。该数据集主要面向复制移动篡改，复制区域与原始背景之间往往具有较强的纹理和结构相似性，真实相似区域也容易形成干扰。EARG-Net的重建残差机制能够直接放大复制边界附近的细微不一致，Co-Transformers通过宏观语义与微观痕迹双分支协同，也有助于处理复制区域与上下文之间的关系。因此，这两类方法在固定阈值F1分数上具有优势。本文方法的F1分数为0.603，低于EARG-Net，但与FRD-Net接近，并在已报告IoU的模型中取得最高IoU（0.543），说明本文方法虽然对部分弱边界复制区域的响应偏保守，但在已检出的篡改区域上具有较好的区域重叠质量和边界控制能力。

综合来看，近期方法与本文方法的侧重点并不相同。EARG-Net更依赖于重建先验与残差放大机制，适用于弱边界、修复型及复制粘贴类篡改样本。FRD-Net强调宏观与微观差异的联合增强，在篡改区域与真实区域差异较为明显的场景中具有优势。Co-Transformers通过双Transformer分支分别建模宏观语义逻辑与微观篡改痕迹，适合处理语义结构与局部伪迹并存的复杂情形。SparseViT、Mesorch和NCNeT则分别从非语义稀疏建模、介观多尺度结构和噪声感知交互角度提升定位性能。相比之下，本文方法并非单独强化某一种线索，而是通过显式解耦和伪迹先验引导融合机制，在细粒度伪迹感知与粗粒度结构建模之间建立动态协同关系。因此，尽管本文方法在Coverage和NIST16的固定阈值F1分数上低于EARG-Net，但在平均F1分数、平均IoU和平均AUC上仍表现出更均衡的跨数据集泛化能力。

2.4　可视化结果分析

为更直观地评估各方法在篡改区域检测与定位任务中的表现，本文选取多个具有代表性的测试样例，对不同模型的预测结果进行了可视化对比，不同模型在数据集上的可视化结果如图2所示。实验表明，本文方法在不同数据集与不同场景下的预测结果均与真实掩码更为接近。相较于其他方法，

当图像存在复杂背景、目标尺度较小或篡改区域形状不规则时，本文方法能够有效抑制背景噪声的干扰，减少零散误检，并保持预测区域的完整性与连续性。特别地，在部分对比方法出现边界偏移或区域缺失的样例中，本文方法仍能较为准确地刻画篡改区域的整体轮廓，在边界贴合度与区域一致性方面均展现出更优的视觉效果。由此表明，本文方法不仅能够有效检出篡改区域，在精细定位方面也具有良好的稳定性。

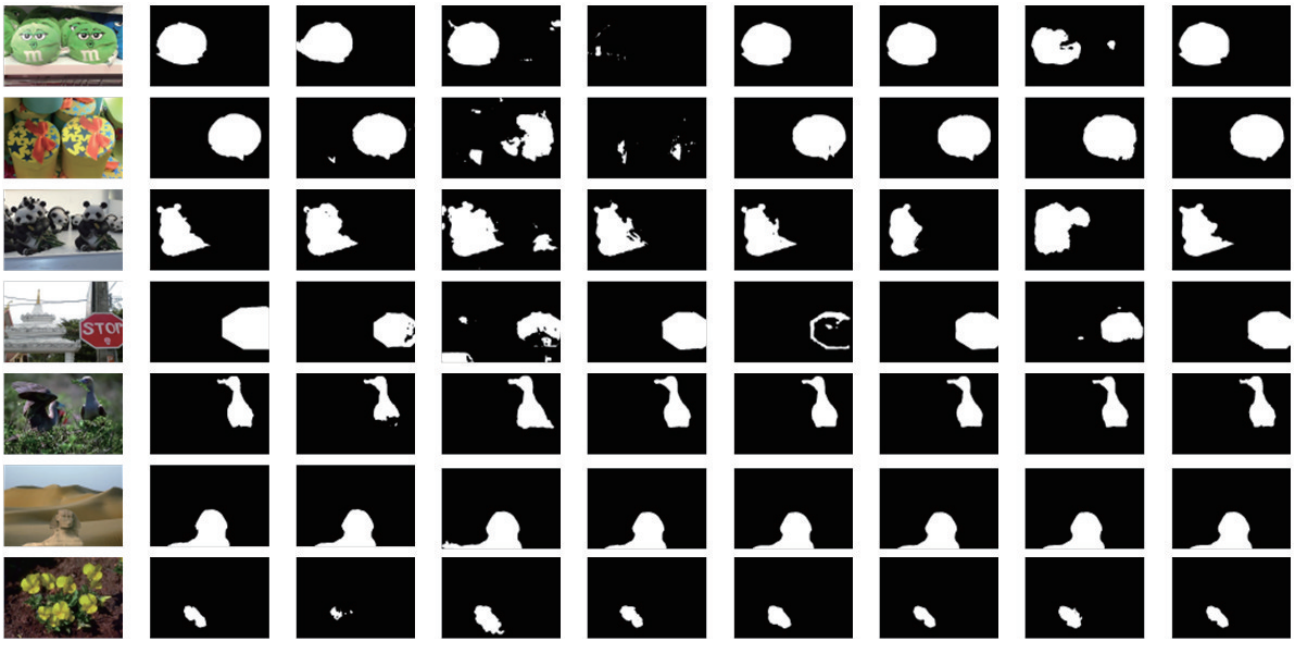
结合图 2 与表 1 可以看出，可视化结果与 F1 分数和 IoU 指标的变化趋势基本一致。这表明本文所构建的多粒度表征机制能够较好地兼顾局部细节刻画与全局结构建模，有助于增强模型对边界邻域、纹理异常以及微弱篡改痕迹的感知能力。与此同时，显式解耦模块进一步明确了不同粒度特征之间的分工，门控自适应融合模块则能够依据空间位置差异动态调节融合权重，使模型在边界区域保留更丰富的细节信息，在非篡改背景区域抑制冗余响应。因而，本文方法在可视化结果中表现出更清晰的边界、更完整的区域结构以及更少的误检，这也从另一个角度说明了所提方法的有效性。

2.5　消融实验

为评估各关键模块对模型性能的影响，本文在 CASIA_V1 数据集上进行了消融实验。本文实验对照模型为去除本文所提出关键模块后的基础篡改定位网络。具体而言，该模型保留与完整模型一致的输入设置、主干特征提取框架、基本解码结构和主定位损失，但不引入多粒度表征构建、显式解耦约束以及伪迹先验引导的门控自适应融合模块，即仅依赖常规编码与解码路径完成像素级篡改区域预测。在此基础上，本文依次加入多粒度特征构建、显式解耦模块及不同融合策略，构建多个模型变体，以分析各模块和融合策略的实际作用。所有实验均保持训练数据、训练参数与评价方式一致。消融实验结果如表 3 所示，可视化结果如图 3 所示。

表 3　　消融实验结果

实验设置	F1 分数	IoU	AUC
对照	0.752	0.658	0.962
多粒度	0.771	0.673	0.974
多粒度+显式解耦	0.811	0.721	0.980
多粒度+显式解耦+相加融合	0.813	0.748	0.976
多粒度+显式解耦+拼接融合	0.449	0.323	0.562
多粒度+显式解耦+无自适应融合	0.824	0.760	0.974
多粒度+显式解耦+自适应融合	0.838	0.780	0.988



(a) 篡改图片　(b) 图片掩码　(c) TruFor　(d) CoDE　(e) EITLNeT　(f) SparseViT　(g) Mesorch　(h) NCNeT　(i) 本文方法

图 2　不同模型在数据集上的可视化结果

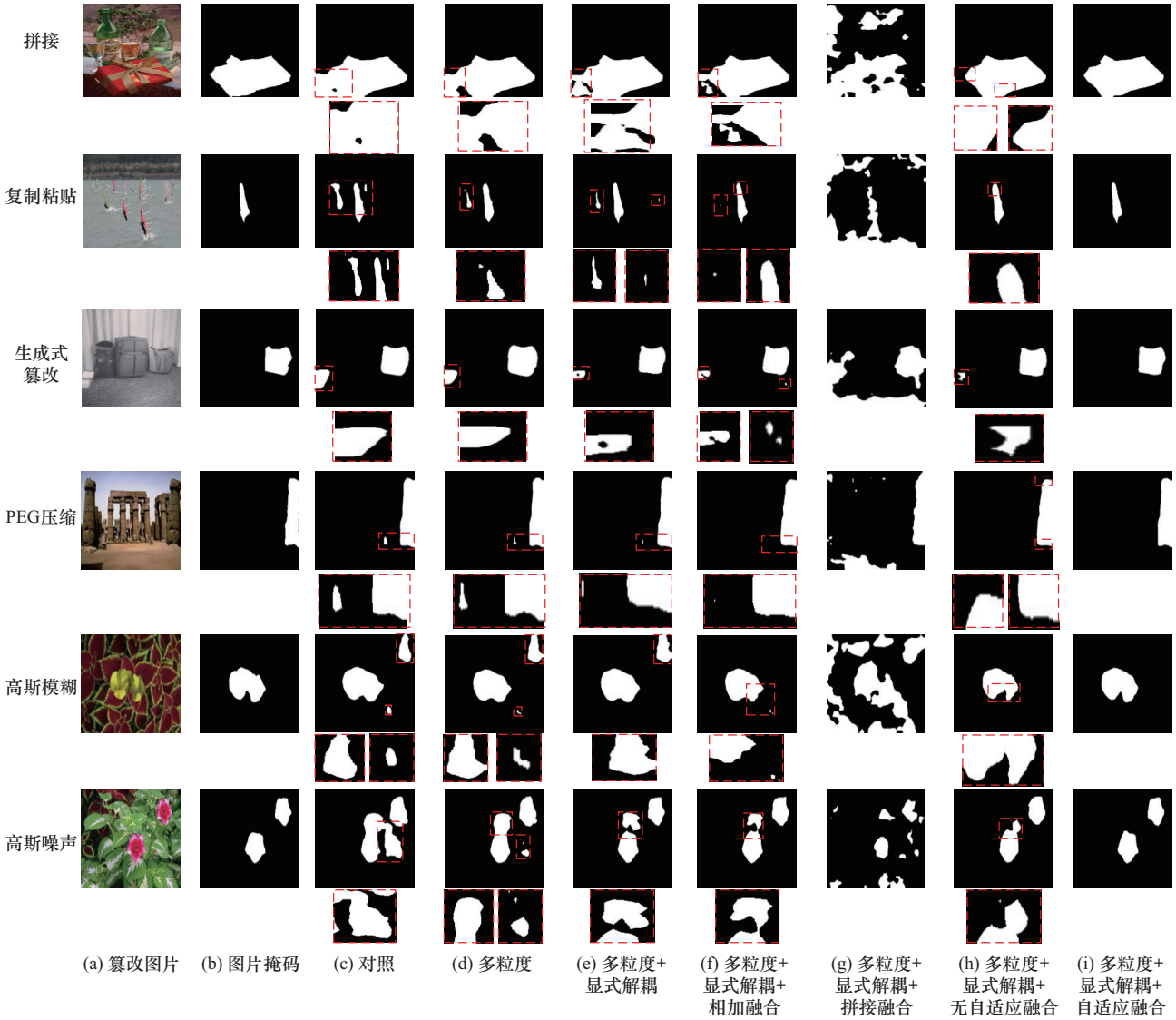


图3　消融实验可视化

由表3可以看出，对照模型的F1分数、IoU和AUC分别为0.752、0.658和0.962，说明常规编码–解码器已具备一定的篡改区域定位能力。引入多粒度特征后，F1分数提升至0.771，IoU提升至0.673，表明细粒度伪迹线索和粗粒度结构信息的联合建模能够改善模型对篡改区域的表达能力。进一步加入显式解耦模块后，F1分数和IoU分别提升至0.811和0.721，较仅使用多粒度特征时提升更为明显。这表明仅引入多粒度特征并不必然带来有效互补。各分支仍在同一像素级监督下优化，不同粒度特征可能共同响应篡改主体或显著内容区域，导致细粒度伪迹线索与粗粒度结构信息之间存在冗余和功能重叠。显式解耦模块通过降低二者之间的冗余相关性，使不同粒度表征形成更清晰的分工，从而进一步提升区域重叠质量和定位精度。

在融合策略方面，逐元素相加融合的F1分数仅由0.811提升至0.813，增益较小，说明简单相加缺少空间选择性，难以根据边界区域、篡改区域内部和复杂背景区域的不同需求动态分配特征贡献。通道拼接融合的F1分数和IoU分别下降至0.449和0.323，性能下降明显，说明保留更多通道并不等同于有效融合。由于细粒度特征包含较多纹理、噪声和边界响应，粗粒度特征包含更多区域结构和上下文信息，直接拼接会增加解码器筛选有效线索的难度，并可能放大无关背景响应。相比之下，无自适应融合将F1分数和IoU提升至0.824和0.760，表明门控机制能够自适应调节局部分支与上下文分支的贡献，比静态融合方式更适合多粒度特征协同。

但无自适应融合仍完全依赖特征自身学习权重，容易受到场景语义、目标外观或背景纹理的影响。引入伪迹先验后，完整模型的 F1 分数、IoU 和 AUC 分别达到 0.838、0.780 和 0.988，均为最优结果。结果说明伪迹先验能够为门控权重提供与噪声异常、压缩不一致和纹理突变相关的任务引导，融合过程稳定聚焦真实篡改线索。

图 3 中虚线框标出了各模型预测差异较明显的局部区域，主要包括篡改边界、小区域漏检、背景误检和区域断裂等位置。从图 3 可以看出，对照模型尽管未引入多粒度解耦与自适应融合，但仍保留了像素级监督下的编码–解码定位结构，能够学习篡改区域与真实区域之间较显著的颜色、纹理、边界和局部统计差异。因此，对于篡改痕迹较明显或主体区域较大的样例，对照模型可以大致检出篡改区域。然而，该模型缺少细粒度伪迹表征与粗粒度结构表征的分工建模，也缺少空间自适应融合机制，在弱边界、小区域和复杂背景干扰下易出现边界偏移、局部漏检或区域断裂等问题。图 3 第 1 行展示了拼接篡改的情况，不同来源区域之间往往存在光照、压缩和边缘过渡差异，完整模型通过联合利用局部异常信息和区域结构信息，使预测结果与真实篡改区域更加贴合。如图 3 第 2 行所示，复制粘贴篡改区域与周围背景在纹理和结构上较为相似，对照模型在边界位置容易出现响应不连续，而完整模型能够更好地保持预测区域的连贯性。对于图 3 第 3 行的生成式篡改，由于篡改区域与周围内容在纹理和语义上具有较强连续性，局部边界特征并不明显，单纯依赖细粒度伪迹容易出现漏检或区域破碎。引入自适应融合后，模型能够适当增强上下文结构约束，从而改善预测区域的完整性和连续性。

为进一步观察各模块在后处理干扰下的响应特性，本文分别对部分样本施加 JPEG 压缩、高斯模糊及高斯噪声扰动。不同扰动下各模型的鲁棒性比较如图 4 所示，随着 JPEG 质量因子降低、模糊程度加剧以及噪声标准差增大，模型定位精度整体呈下降趋势。结合图 4 中的精度变化，本文选取 JPEG 质量因子 80、高斯模糊核大小 7 且标准差 1.5、高斯噪声标准差 7，作为后续消融可视化实验的扰动参数。在该参数组合下，模型定位精度相较于其他设置呈现出明显变化，表明篡改痕迹受到有效削弱。同时，模型整体性能尚未出现严重退化，图像主体结构仍得以较好保持，因此，该设置适于考察各模块在常见质量退化情形下的定位稳定性。在该类后处理场景下，对照模型更容易出现边界模糊、小区域漏检或背景误检。多粒度特征能够同时补充局部异常和区域结构信息，显式解耦进一步减少不同粒度特征之间的相互干扰，而自适应融合则根据伪迹响应强弱动态调整局部细节和上下文结构的权重，使模型在该类后处理干扰下仍能保持较好的区域完整性。此外，图 3 中各模块带来的提升不总是表现为对主体篡改区域的大幅重新定位。在部分样例中，对照模型已经能够检出主要篡改区域，但其预测结果仍存在边界偏移、局部断裂、小范围漏检或背景误检等问题。这些误差虽然在视觉上属于局部形态差异，但会直接影响像素级精确率、召回率以及预测掩码与真实掩码之间的重叠质量。引入多粒度表征、显式解耦和自适应融合后，模型能够进一步修正上述局部误差，使预测区域在边界贴合、区域连续性和误检抑制方面更加稳定。这与表 3 中

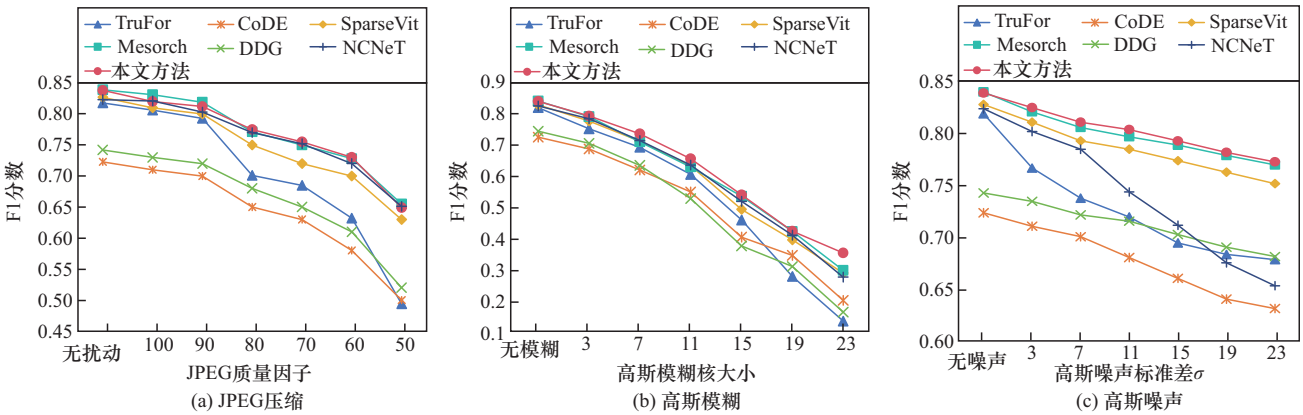


图 4　不同扰动下各模型的鲁棒性比较

的定量结果总体一致，说明本文模块主要改善的是复杂条件下的精细定位质量和预测稳定性。

综合定量结果和可视化分析可以看出，多粒度表征构建、显式解耦模块和自适应融合均对模型性能具有积极作用。其中，显式解耦模块有效降低了不同粒度表征之间的冗余耦合，是提升定位性能的重要因素。自适应融合则进一步说明，多粒度特征在解码阶段需要根据空间位置和伪迹响应进行自适应协同，而不能简单采用相加或拼接。最终，完整模型在F1分数、IoU和AUC 3个指标上均取得最优结果，验证了本文整体设计的有效性。

2.6 鲁棒性实验

为进一步验证所提方法在常见后处理攻击下的稳定性，本文在CASIA_V1数据集上进行了鲁棒性实验，分别测试了JPEG压缩、高斯模糊和高斯噪声3类典型扰动，并与TruFor、CoDE、SparseViT、DDG、Mesorch和NCNeT方法进行了比较，评价指标为像素级F1分数。其中，JPEG压缩通过改变质量因子模拟不同程度的压缩失真，质量因子越小表示压缩越强。高斯模糊通过改变滤波核大小模拟边界平滑退化，核大小越大表示模糊越明显。高斯噪声通过改变噪声标准差模拟随机噪声干扰，标准差越大表示噪声越强。不同扰动下各模型的鲁棒性比较如图4所示。实验结果表明，随着扰动强度增加，各方法性能均呈现不同程度下降，但本文方法整体上始终保持较强竞争力。如图4(a)所示，在JPEG压缩条件下，本文方法在无压缩及轻中度压缩场景中保持较高性能，仅在质量因子降至50时略低于Mesorch。如图4(b)和图4(c)所示，在高斯模糊条件下，随着滤波核增大，边界和纹理细节被逐步平滑，各方法性能下降明显，但本文方法仍保持较好的稳定性。高斯噪声条件下，随机噪声会干扰原有噪声统计分布，本文方法整体仍处于对比方法前列。综合来看，本文方法在3类常见后处理条件下均保持较优的性能，说明该方法不仅在无扰动场景下具有较好的篡改检测与定位能力，而且在压缩失真、边界平滑和噪声污染等干扰下仍具备较好的鲁棒性，这也进一步验证了所提多粒度表征解耦与协同机制在复杂场景下的有效性。

2.7 讨论与局限性

实验结果表明，本文方法在跨数据集测试、消融实验和常见后处理扰动下整体表现较为稳定，表明多粒度表征解耦与协同融合能够提升模型的定位精度和泛化能力。本文方法的显式解耦模块主要面向细粒度伪迹表征与粗粒度结构表征之间的冗余抑制，通过归一化嵌入空间中的去相关约束促进二者形成更清晰的功能分工。该模块重点在于增强局部伪迹信息与区域结构信息之间的互补性，使模型在多数跨域场景中保持较稳定的定位结果。

从图2和图3可以发现，本文方法的增益更多体现在边界贴合、小范围漏检修正、局部误检抑制和区域形态优化等方面。当对照模型已经能够较好检出主体篡改区域时，后续模块未必会显著改变整体区域的预测结果，而更多表现为预测结果精细程度与稳定性的改善。从跨数据集结果看，本文方法在平均性能上具有较好的综合表现，但在Coverage和NIST16等弱边界、复杂编辑样本较多的数据集上，固定阈值F1分数仍存在提升空间。后续可进一步围绕弱边界、强融合篡改和复杂后处理样本优化局部异常建模，使模型在保持区域结构稳定性的同时，增强对细微篡改线索的响应能力。

3 结束语

本文提出了一种基于多粒度表征解耦与协同的图像篡改检测与定位方法，用于提升复杂场景下篡改区域的检测与精细化定位能力。该方法在统一网络中同时建模细粒度异常表征和粗粒度结构表征，使模型不仅能够关注边界断裂、纹理突变等局部伪迹线索，也能把握篡改区域的结构连续性和整体一致性。因此，在面对复杂边界、小尺寸篡改区域的情况时，模型可以更稳定地完成篡改区域定位。

相较于已有方法中直接堆叠或简单融合多尺度特征的方式，本文方法主要在以下3个方面进行改进。首先，通过多粒度表征建模，使局部伪迹感知与区域结构建模形成协同关系，缓解了单一尺度特征难以兼顾边界细节和区域完整性的问题。其次，引入显式解耦约束，减少了不同粒度表征之间的信息冗余和功能耦合，使细粒度异常表征与粗粒度结构表征能够承担更加明确的作用。最后，设计了伪迹先验引导的门控自适应融合模块，根据不同空间位置的伪迹响应和结构需求动态调整特征融合权重，从而提升了模型在跨数据集场景下的适应能力、定位精度和鲁棒性。

参考文献:

[1] Farid H. Image forgery detection[J]. IEEE Signal Processing Magazine, 2009, 26(2): 16-25.

[2] Popescu A C, Farid H. Exposing digital forgeries by detecting traces of resampling[J]. IEEE Transactions on Signal Processing, 2005, 53(2): 758-767.

[3] Fridrich J, Soukal D, Lukas J. Detection of copy-move forgery in digital images[C]//Proceedings of the Digital Forensic Research Workshop (DFRWS). Piscataway: IEEE Press, 2003: 55-61.

[4] 王珠珠. 基于U型检测网络的图像篡改检测算法[J]. 通信学报, 2019, 40(4): 171-178.
Wang Z Z. Image forgery detection algorithm based on U-shaped detection network[J]. Journal on Communications, 2019, 40(4): 171-178.

[5] Bappy J H, Roy-Chowdhury A K, Bunk J, et al. Exploiting spatial structure for localizing manipulated image regions[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 4980-4989.

[6] Zhou P, Han X T, Morariu V I, et al. Learning rich features for image manipulation detection[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 1053-1061.

[7] Wu Y, Abd Almageed W, Natarajan P. ManTra-net: manipulation tracing network for detection and localization of image forgeries with anomalous features[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 9535-9544.

[8] Zhou P, Chen B C, Han X T, et al. Generate, segment, and refine: towards generic manipulation segmentation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 13058-13065.

[9] Chen X R, Dong C B, Ji J Q, et al. Image manipulation detection by multi-view multi-scale supervision[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 14165-14173.

[10] 朱叶, 余宜林, 郭迎春. HRDA-Net: 面向真实场景的图像多篡改检测与定位算法[J]. 通信学报, 2022, 43(1): 217-226.
Zhu Y, Yu Y L, Guo Y C. HRDA-Net: image multiple manipulation detection and location algorithm in real scene[J]. Journal on Communications, 2022, 43(1): 217-226.

[11] Zhou J Z, Ma X C, Du X, et al. Pre-training-free image manipulation localization through non-mutually exclusive contrastive learning[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2023: 22289-22299.

[12] Kwon M J, Yu I J, Nam S H, et al. CAT-net: compression artifact tracing network for detection and localization of image splicing[C]//Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE Press, 2021: 375-384.

[13] Wang J K, Wu Z X, Chen J J, et al. ObjectFormer for image manipulation detection and localization[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 2354-2363.

[14] Liu X H, Liu Y J, Chen J, et al. PSCC-net: progressive spatio-channel correlation network for image manipulation detection and localization [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(11): 7505-7517.

[15] 谭舜泉, 廖桂樱, 彭荣煊, 等. 面向强后处理场景的图像篡改定位模型[J]. 通信学报, 2024, 45(4): 146-159.
Tan S Q, Liao G Y, Peng R X, et al. Image tampering localization model for intensive post-processing scenarios[J]. Journal on Communications, 2024, 45(4): 146-159.

[16] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale[C]//International Conference on Learning Representations. Vancouver: ICLR, 2021: 1-22.

[17] Cozzolino D, Verdoliva L. Noiseprint: a CNN-based camera model fingerprint[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 144-159.

[18] Bianchi T, Piva A. Image forgery localization via block-grained analysis of JPEG artifacts[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 1003-1017.

[19] Child R, Gray S, Radford A, et al. Generating long sequences with sparse transformers[PP]. V1. (2019-04-23) [2026-04-22]. arXiv: arXiv.1904.10509.

[20] Belghazi M I, Baratin A, Rajeshwar S, et al. Mutual information neural estimation[C]//Proceedings of the 35th International Conference on Machine Learning. New York: PMLR, 2018: 531-540.

[21] Li X, Wang W H, Hu X L, et al. Selective kernel networks[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 510-519.

[22] 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.

[23] Dong J, Wang W, Tan T N. CASIA image tampering detection evaluation database[C]//Proceedings of the 2013 IEEE China Summit and International Conference on Signal and Information Processing. Piscataway: IEEE Press, 2013: 422-426.

[24] Novozamsky A, Mahdian B, Saic S. IMD2020: a large-scale annotated dataset tailored for detecting manipulated images[C]//Proceedings of the 2020 IEEE Winter Applications of Computer Vision Workshops (WACVW). Piscataway: IEEE Press, 2020: 71-80.

[25] Kniaz V V, Knyaz V, Remondino F. The point where reality meets fantasy: mixed adversarial generators for image splice detection[C]//Advances in Neural Information Processing Systems 32. Massachusetts: MIT Press, 2019: 215-226.

[26] Wen B H, Zhu Y, Subramanian R, et al. COVERAGE: a novel database for copy-move forgery detection[C]//Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE Press, 2016: 161-165.

[27] Ng T T, Chang S F. A model for image splicing[C]//Proceedings of the 2004 International Conference on Image Processing, 2004. ICIP '04. Piscataway: IEEE Press, 2004: 1169-1172.

[28] Guan H Y, Kozak M, Robertson E, et al. MFC datasets: large-scale benchmark datasets for media forensic challenge evaluation[C]//Pro-

ceedings of the 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). Piscataway: IEEE Press, 2019: 63-72.

[29] Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library[C]//Advances in Neural Information Processing Systems 32. Massachusetts: MIT Press, 2019: 8024-8035.

[30] Loshchilov I, Hutter F. Decoupled weight decay regularization[C]//International Conference on Learning Representations. Vancouver: ICLR, 2019: 1-18.

[31] Guillaro F, Cozzolino D, Sud A, et al. TruFor: leveraging all-round clues for trustworthy image forgery detection and localization[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 20606-20615.

[32] Peng R X, Tan S Q, Mo X B, et al. Employing reinforcement learning to construct a decision-making environment for image forgery localization[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 4820-4834.

[33] Guo K, Zhu H C, Cao G. Effective image tampering localization via enhanced transformer and co-attention fusion[C]//Proceedings of the ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2024: 4895-4899.

[34] Su L, Ma X C, Zhu X K, et al. Can we get rid of handcrafted feature extractors? SparseViT: nonsemantics-centered, parameter-efficient image manipulation localization through spare-coding transformer[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(7): 7024-7032.

[35] Huang Y H, Luo W Q, Cao X C, et al. A forensic framework with diverse data generation for generalizable forgery localization[J]. IEEE Transactions on Information Forensics and Security, 2025, 20: 9732-9745.

[36] Zhu X K, Ma X C, Su L, et al. Mesoscopic insights: orchestrating multi-scale & hybrid architecture for image manipulation localization[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(10): 11022-11030.

[37] Zhang H S, Su T H, Liu Z, et al. Noise-aware cross attention for image manipulation localization[J]. Pattern Recognition, 2026, 176: 113164.

[38] Chen S H, Zhao Y B, Wang T Y, et al. Amplifying discrepancies: exploiting macro and micro inconsistencies for image manipulation localization[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2026, 40(42): 35357-35365.

[39] Zhang J W, Feng W B, Wang S W, et al. Collaborative transformers with multi-level forensic attention for image manipulation localization[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2026, 40(15): 12556-12563.

[40] Yu Y P, Shi Z X, Zhao H Q, et al. EARG-net: edge-aware reconstruction-guided network for image manipulation detection and localization[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2026, 40(14): 12178-12186.

[作者简介]



袁程胜（1989– ），男，南京信息工程大学副教授，主要研究方向为人工智能安全、信息隐藏。



曹富贵（2000– ），男，南京信息工程大学硕士生，主要研究方向为多媒体内容安全、图像取证。



王一力（1989– ），男，南京信息工程大学讲师，主要研究方向为信息安全、深度学习。



曹燚（1994– ），男，无锡学院副教授，主要研究方向为信息隐藏。



刘庆程（1987– ），男，南京信息工程大学博士生，主要研究方向为多媒体内容安全、电力大数据挖掘与分析。



付章杰（1983– ），男，南京信息工程大学教授，主要研究方向为信息安全。