E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS


**中图法分类号**:TP391;TP18　　**文献标识码**: A　　**文章编号**: 1006-8961(2026)04-1090-18

**论文引用格式**:Liu F Y, Zhang Y J and Wu F. 2026. Multiview vision-language interaction for multimodal media content manipulation detection and localization. Journal of Image and Graphics, 31(4):1090-1107(刘凤阳, 张玉金, 吴飞. 2026. 多视角视觉—语言交互的多模态媒体内容篡改检测与定位. 中国图象图形学报, 31(4):1090-1107)[DOI:10.11834/jig.250414]

# 多视角视觉—语言交互的多模态媒体内容篡改检测与定位

刘凤阳,张玉金*,吴飞

上海工程技术大学电子电气工程学院,上海　201620

**摘　要**:**目的**　错误信息的传播已成为数字时代亟待解决的重大挑战。随着多媒体技术的快速发展,网络空间中视觉与文本模态相结合的虚假内容呈现泛滥态势。尽管现有研究在多模态媒体篡改检测与定位方面取得了一定进展,但普遍存在跨模态层次化信息交互不足、篡改区域定位精度有限等关键问题。针对上述挑战,提出了一种基于多视角视觉—语言信息交互的篡改检测框架。**方法**　首先,通过全局与局部双视角特征嵌入,构建层次化篡改对比学习机制,实现跨模态细粒度语义对齐,有效捕捉篡改区域的语义不一致性;其次,创新性地设计了伪造感知交互模块,集成多尺度特征提取与频域特征融合策略,显著提升了对不同粒度篡改特征的定位能力;此外,引入跨模态门控融合模块,采用动态权重分配策略优化模态间信息交互,从而增强模型在多模态深度伪造检测及细粒度分类任务中的判别能力。**结果**　实验结果表明,在相同实验环境下,本模型相较基于分层推理的HAMMER(hierarchical multi-modal manipulation reasoning Transformer)框架,在图像深度伪造定位任务中IoU75(intersection over union)指标提升6.41%,文本篡改定位任务的召回率与F1分别提高5.63%和2.01%。与VLP-GF(visual-language pre-training with gate fusion)框架相比,本模型在多模态多任务学习的综合评估中展现出全面性能优势。**结论**　本文提出的多视角视觉—语言信息交互模型相较于其他模型,在多模态深度伪造检测与定位任务中表现出显著优越性,为多媒体内容安全领域提供了新的技术解决方案。

**关键词**:多模态深度伪造检测;视觉—语言交互;特征融合;篡改定位;跨模态交互

# Multiview vision-language interaction for multimodal media content manipulation detection and localization

Liu Fengyang, Zhang Yujin*, Wu Fei

*School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai　201620, China*

**Abstract**: **Objective** The propagation of misinformation has emerged as a significant challenge in the digital age. Traditional multimodal fake news detection research has primarily focused on the binary classification task of content authenticity; for instance, the capabilities for identifying specific types of tampering and localizing tampered regions are lacking. Fake content that combines visual and textual modalities in cyberspace has proliferated rapidly with the rapid advancement of multimedia technologies. Despite some progress in multimodal media tampering detection and localization, existing stud-

**收稿日期**:2025-09-03;**修回日期**:2025-11-04;**预印本日期**:2025-11-11

*通信作者:张玉金　yjzhang@sues.edu.cn

**基金项目**:国家自然科学基金项目(62072057);上海市自然科学基金项目(17ZR1411900);中国高校产学研创新基金项目(2021ZYB01003)

**Supported by**: National Natural Science Foundation of China(62072057); Shanghai Municipal Natural Science Foundation(17ZR1411900); Innovation Fund for Industry-University-Research of Chinese Universities(2021ZYB01003)

ies commonly face critical issues, including insufficient cross-modal hierarchical information interaction and limited accuracy in localizing tampered regions. This study addresses these challenges by proposing a tampering detection framework based on multiview visual-language information interaction. **Method** The proposed method is designed to tackle the key challenges in multimodal deep forgery detection and localization by leveraging multiple perspectives of visual-language interaction. First, a hierarchical tampering contrastive learning mechanism is constructed using global and local dual-view feature embedding. This mechanism achieves fine-grained cross-modal semantic alignment, effectively capturing the semantic inconsistency in tampered regions. The approach innovatively integrates multiple strategies to enhance the model's ability to detect and localize forgeries. Second, a forgery perception interaction module is designed. It incorporates multiscale feature extraction and frequency-domain feature fusion techniques. In this way, the localization capability of tampered features at different granularities is significantly improved. In addition, a cross-modal gating fusion module is introduced to optimize the information interaction between modalities. This approach is achieved through a dynamic weight allocation strategy that enhances the model's discriminative power in multimodal deep forgery detection and fine-grained classification tasks. The proposed method utilizes advanced techniques in deep learning, including transformers and bidirectional encoder representations from transformers (BERT), to achieve fine-grained alignment and tampering detection. Vision transformers (ViTs) are used for image feature extraction, whereas BERT is used for extracting textual features. The embedded features are subjected to tampering contrastive learning from global and local perspectives. Unlike conventional methods, which typically only pull close-matching image-text pairs and push away nonmatching pairs, the proposed approach employs the InfoNCE loss function to push away tampered image-text pairs simultaneously, thereby reinforcing their semantic inconsistency. This method enhances the model's ability to distinguish between tampered and nontampered content, thereby addressing the critical issue of cross-modal alignment in traditional methods. Moreover, a dual-stream cross-modal attention mechanism is utilized to facilitate deep-level information interaction between visual and textual modalities. This mechanism enables the model to capture detailed information from both modalities and improve the overall detection accuracy. In addition to the attention mechanism, the forgery perception interaction module and cross-modal gating fusion module further refine the model's ability to detect subtle discrepancies in content across different modalities. These modules collectively improve the detection of fine-grained tampering, which is crucial for accurately localizing tampered regions in multimodal content. A multitask learning framework is employed to handle various downstream tasks simultaneously, including bounding box regression, binary classification, multilabel classification, and token-level tampering detection. The multitask learning module consists of a set of lightweight multilayer perceptrons (MLPs) designed to handle these tasks efficiently. The model can simultaneously learn from various types of labeled data and generalize across different types of tasks by employing a multitask learning approach, thereby providing a comprehensive solution for multimodal deep forgery detection. Specific Implementation: The proposed framework is implemented on a high-performance computing environment using the DGM$^4$ dataset. The model is trained for 50 epochs on an NVIDIA RTX 4090 GPU, leveraging the PyTorch framework. The dataset used for training contains various multimodal content, including images and corresponding text, which is essential for evaluating the model's effectiveness in detecting and localizing tampered content across different modalities. ViT is used for extracting image features, which allows the model to process visual data efficiently and capture detailed spatial information. BERT, a powerful transformer-based model for natural language processing, is used for text feature extraction. Then, these features are embedded into a shared space, where tampering contrastive learning is applied. The contrastive learning mechanism ensures that the embeddings of tampered image-text pairs are pushed away from the embeddings of authentic image-text pairs, thereby reinforcing the semantic inconsistencies in tampered regions. This fine-grained contrastive learning approach allows the model to identify subtle discrepancies in both the image and text modalities, thereby improving the detection of tampered regions. In terms of information interaction, the dual-stream cross-modal attention mechanism is employed to allow for deep interaction between the visual and textual features. This mechanism facilitates the exchange of information between the two modalities, thereby enabling the model to capture complex relationships between the image and the text. The forgery perception interaction module further enhances the model's ability to detect subtle tampering by integrating multiscale features and fusing frequency-domain information. This approach enables the model to capture tampering features at different scales and granularities, thereby improving its

ability to localize tampered regions effectively. The cross-modal gating fusion module is another crucial component of the proposed framework. It optimizes the interaction between modalities by using a dynamic weight allocation strategy. This strategy ensures that the most relevant features from each modality are given high importance during the decision-making process, thereby improving the model's discriminative power. This module is particularly effective in scenarios where the tampered content is subtle or involves complex interactions between visual and textual elements. The multitask learning module, consisting of lightweight MLPs, is designed to support various downstream tasks. These tasks include bounding box regression for localizing tampered regions, binary classification for distinguishing tampered and authentic content, multilabel classification for handling multiple types of tampering, and token-level tampering detection for identifying tampered tokens in text. The multitask framework allows the model to learn from multiple sources of supervision and generalize across different tasks, thereby improving its overall performance. **Result** Experimental results demonstrate that the proposed model outperforms existing approaches, such as the hierarchical multimodal manipulation reasoning transformer framework based on hierarchical reasoning, in multimodal deep forgery localization tasks. In the image deep forgery localization task, the model achieves a 6.41% improvement in the intersection over union at a threshold of 75% (IoU75) metric, thereby indicating a significant enhancement in the precision of localized tampered regions. In text tampering localization tasks, the model improves the recall and F1 scores by 5.63% and 2.01%, respectively, thereby demonstrating its superior ability to detect tampered text content. These improvements are a direct result of the fine-grained alignment and deep interaction between visual and textual features enabled by the proposed framework. Compared with the visual-language pretraining with a gate fusion framework, the proposed model exhibits a comprehensive performance advantage in the evaluation of multimodal multitask learning. The model's ability to handle various tasks simultaneously, coupled with its robust performance in image and text modalities, makes it a highly effective solution for multimodal deep forgery detection and localization. **Conclusion** The multiview visual-language information interaction model proposed in this paper exhibits significant superiority over other models in multimodal deep forgery detection and localization tasks, thereby providing a novel technical solution for the multimedia content security field.

**Key words**: multimodal deepfake detection; visual-language interaction; feature fusion; manipulation grounding; cross-modality interaction

## 0 引　言

近年来,深度学习的快速发展极大地提升了通过深度生成模型和大型语言模型(Devlin 等,2019;Radford 等,2019)合成媒体内容的能力。借助先进的生成对抗网络(Goodfellow 等,2014),这些技术能够生成高度逼真但完全人工合成的多媒体内容,包括深度伪造的人脸图像、伪造视频以及虚构文本。然而,深度伪造技术的滥用现象日益普遍,其模型致力于制造看似真实实则虚构的欺骗性虚假信息。此类信息的传播会带来重大风险,包括误导公众、制造混乱,并可能侵害个人声誉与隐私。

为减轻深度伪造的负面影响,研究者正积极开发深度伪造检测技术,旨在有效识别并应对误导信息构成的普遍威胁。根据被伪造内容类型,深度伪造检测领域可分为两大方向:单模态方法与多模态方法。单模态方法专注于处理某一特定模态(如视觉或文本内容),采用针对该模态的特定算法进行伪造检测;而多模态方法则融合来自不同模态的数据,通过跨模态特征融合与语义一致性分析,提升检测性能。多项研究致力于发展先进的单模态深度伪造检测方法(张晶 等,2025),以应对伪造媒体内容传播所带来的挑战。Monu 和 Dhanakshirur(2024)提出了一种启发式多阶段学习框架,该框架在包含眼部遮挡的合成数据集上对模型进行预训练,从而引导其关注眼部以外的面部特征。同时,为了解决类别不平衡问题,研究者还引入了加权损失函数。Sun 等人(2024)提出了 DiffusionFake,这是一种可插拔式框架,利用预训练扩散模型 Stable Diffusion,引导伪造检测器学习源人脸与目标人脸之间的可分离特征。此外,Galdi 等人(2024)提出了 2D-Malafide 攻击方法,该方法通过卷积噪声扰动对深度伪造检测系统进行对抗攻击,专门针对其对图像特定依赖和泛化能力的局限性。现有单模态深度伪造检测方法在不同维度上取得了一定进展,但它们普遍依赖于

单一模态的特征表示，难以应对跨模态场景下更为复杂的语义不一致问题。例如，图像与文本的篡改往往是协同出现的，单模态方法无法有效捕捉两种模态之间潜在的逻辑冲突和语义错配。因此，研究者开始转向多模态深度伪造检测（王诗雨 等，2025），期望通过融合视觉与语言等多源信息，实现更鲁棒、更细粒度的伪造检测。Salvi 等人（2023）提出了一种用于分析音频—视觉特征中时间不一致性的全新框架，其方法从不同的单模态数据集中提取合成音频与视觉特征，并在训练与推理阶段进行融合，以应对完全对齐的多模态数据集稀缺的问题。Zou 等人（2024）针对音频与视频模态间的一致性问题，引入了跨模态与模态内正则化技术，并将 Transformer模块整合进音视频特征提取过程中，从而增强了模态间对应关系，显著提升了检测精度，并在多个数据集上展现出良好的泛化能力。Amoroso 等人（2024）探讨了通过联合分析图像与文本描述检测伪造内容的可行性，提出了一种对比性解耦方法，使模型能够分离低层感知特征与高层语义特征，从而提升伪造检测精度，并减弱由生成模型引入的低层伪影的干扰。总体来看，现有多模态深度伪造检测方法主要集中在不同模态特征的一致性建模上。尽管这些工作在不同方向上推动了多模态伪造检测的发展，但它们大多仍聚焦于“是否存在伪造”的判别任务，缺乏对伪造内容具体位置和形式的精细化刻画。而在真实应用场景中，伪造往往不仅涉及“有/无”的判别，还需要明确指出篡改的区域或对应的文本片段，以实现更具可解释性和可操作性的检测。因此，实现多模态深度伪造的检测与定位成为亟需解决的问题。

为弥补这一不足，Shao 等人（2023）首次构建了用于多模态篡改内容检测与定位的数据集——DGM$^4$（detecting and grounding multi-modal media manipulation），旨在识别并定位来自以人为中心的新闻图文对中的篡改内容。为应对此任务，作者提出 HAMMER（hierarchical multi-modal manipulation reasoning Transformer）框架，该框架采用分层推理机制：浅层推理用于定位图像中的人脸篡改区域，深层推理则处理其余子任务，但分层推理框架信息交互不足，在复杂跨模态场景中容易受到任务间相互干扰的影响。Zhao 等人（2024）提出了集中式推理与统一重建（concentrated reasoning and unified reconstruction，CrUr）模型，通过掩码信号建模实现跨模态信息的重建，其集中式建模方式在处理细粒度语义错配时表现有限。Liu 等人（2025）提出了统一频域辅助 Transformer 框架（unified frequency-assisted Transformer，UFAFormer），该框架通过引入频域信息以及统一解码器结构，提升了多模态篡改检测与定位性能，但其伪造感知仍然主要停留在模态间的线性融合，难以充分建模跨模态的非线性关联。由此可见，这些方法在框架构建、推理机制或融合策略上各有侧重，却普遍存在对复杂跨模态关系建模不足、融合异构特征有限等问题。

针对这些不足，本文提出了一种新型框架——多视角视觉—语言信息交互模型。该模型从全局与局部两个视角出发，实现视觉与语言特征在任务驱动下的隐式对齐，从而在特征层面更充分地建模跨模态关系。通过多视角的信息交互机制，模型能够有效融合异构模态特征，增强对复杂篡改模式的捕捉能力。在上述总体框架下，本文设计了两个核心技术模块以进一步提升模型性能。首先，引入伪造感知交互模块（forgery-aware interaction module，FAIM）。该模块通过多尺度 Transformer 结构在不同空间分辨率上聚合图像特征，并结合频域分支提取的伪造痕迹特征，利用跨域注意力机制将 RGB 域与频率域信息进行互补融合，从而显著增强模型对局部篡改区域的敏感性和定位精度。这一设计与 UFAFormer（unified frequency-assisted transformer）（Liu 等，2025）基于频域信息的统一解码策略不同：UFAFormer 主要在统一解码阶段利用频域特征增强整体表征，而 FAIM 则在编码与交互阶段就深度对齐空间域与频率域信号，能够更有效地捕捉微小、隐蔽的篡改痕迹。其次，提出跨模态门控融合模块（cross-modal gated fusion module，CGFM）。该模块通过引入动态门控因子，在视觉特征与文本特征融合时自适应地分配模态权重，避免了单一模态在判别过程中的主导性偏差。与 CrUr（Zhao 等，2024）集中式推理与统一重建方法相比，CGFM 并不依赖统一的重建机制，而是通过判别式的任务驱动学习直接优化跨模态交互，从而能够在跨任务学习中保持更强的鲁棒性与可解释性。因此，FAIM 与 CGFM 的引入不仅弥补了现有方法在频域利用与跨模态交互方面的不足，也进一步凸显了本文框架的创新性与实用价值。

本文主要贡献总结如下：1）提出了一种基于局

部视角的篡改对比学习机制。该机制与双流融合架构自然契合，在图像和文本模态中同时存在篡改的场景下，能够实现更精细的语义对齐，有效捕捉篡改区域的细粒度不一致性。2）针对异构特征融合与篡改区域精确定位的问题，引入伪造感知交互模块（FAIM）和跨模态门控融合模块（CGFM），在多任务设置下显著提升了检测与定位的性能。3）实验结果表明，本文提出的多视角视觉—语言信息交互模型具有较强的竞争力和有效性。

## 1 本文模型

### 1.1 模型整体架构

给定一幅描绘特定场景中人物的图像及其对应的文本描述，构成一个图文对。所提出模型的目标是对该图文对进行多层次的分析与检测，包括以下任务：1）识别图文对中被篡改或伪造的内容；2）检测篡改的类型，包括图像内容的变更或文本描述的修改；3）精确定位图像中的篡改区域；4）准确定位并标注文本中被篡改的词语或短语。通过实现上述目标，本模型为多模态信息验证提供了坚实的技术支撑，并为虚假信息的检测与防控提供了一种具有实用价值的解决方案。

本文提出的模型架构如图1所示，主要由3个核心模块构成：多模态特征提取与对比学习模块、多模态特征融合模块和多任务学习模块。三者逐层递进，前两者负责表征与交互建模，第3个模块则在此基础上完成具体任务。

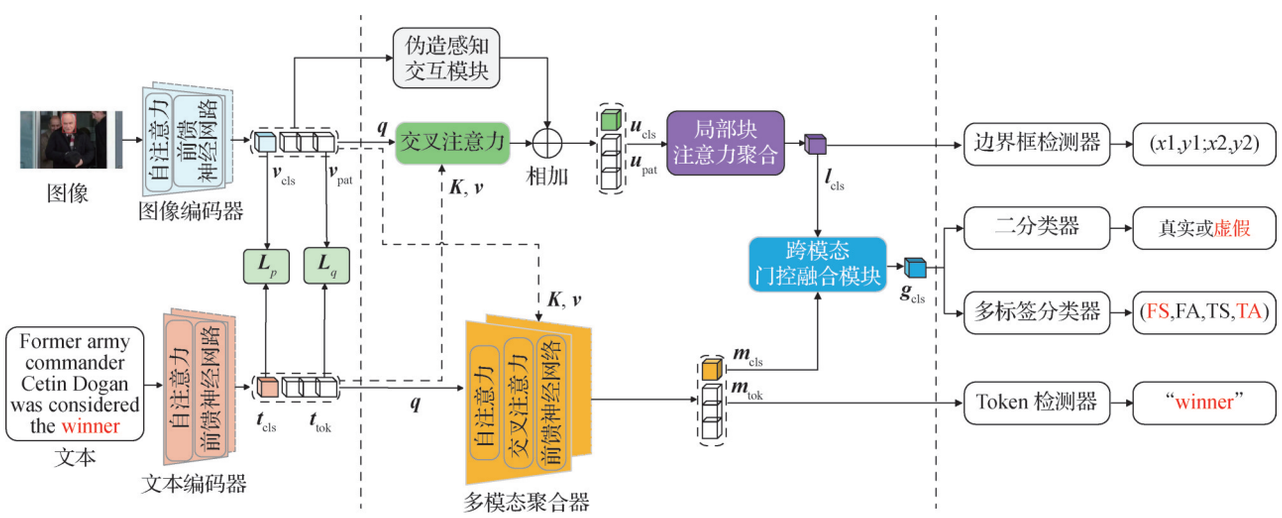



图1 本文模型整体架构

Fig. 1 Overall architecture of the proposed model

((a) multimodal feature extraction and contrastive learning; (b) multimodal feature fusion; (c) multi-task learning)

1）多模态特征提取与对比学习模块。首先，模型需要从图像与文本中提取语义表征。本文采用ViT-B/16（Dosovitskiy等，2021）的12层结构作为图像编码器，以及BERT-base（bidirectional encoder representations from transformers-base version）（Devlin等，2019）的前6层作为文本编码器。在此基础上，为了增强模型捕捉模态之间存在的复杂、非线性关系的能力，从全局与局部两个视角出发，通过对比学习机制实现视觉与语言特征在任务驱动下的隐式对齐。这一阶段不仅提供了模态内的稳健表示，也为后续的跨模态特征融合奠定基础。

2）多模态特征融合模块。在获得初步的模态特征后，模型进一步通过双流跨模态注意力机制进行深度交互。该模块一方面保持图像与文本特征的独立性和内部一致性，为了检测伪造图像中隐藏的细微篡改痕迹，另一方面借助伪造感知交互模块整合多层级视觉线索，以捕捉伪造图像中的细微痕迹。同时，引入局部块注意力聚合（local patch attention aggregation，LPAA）模型提升区域定位能力，并通过跨模态门控融合模块将视觉与语言信息动态整合。此阶段生成的跨模态联合表征直接为下游任务提供输入。

3)多任务学习模块。在前两个模块提供的高质量联合特征基础上,多任务学习模块利用一组轻量级多层感知机(multilayer perceptron,MLP),同时支持二分类、多标签分类、边界框回归和词元级篡改检测。通过这种多任务协同优化,模型不仅能够判别篡改与否,还能在视觉和文本两个层面精确定位具体的篡改内容。

通过各组件间的协同交互,3 个模块形成了一个自底向上的闭环流程:从模态内表征,到跨模态融合,再到任务层面的多维度检测与定位,确保了架构的整体连贯性与功能互补性。模型能够实现鲁棒且细粒度的多模态篡改检测与定位,从而为应对复杂虚假信息场景提供坚实基础。

### 1.2 特征提取

给定一个图文对输入$(\boldsymbol{I},T)$,其中$\boldsymbol{I}$表示图像,$T$表示与该图像相关联的文本描述。视觉Transformer(vision Transformer,ViT)用于图像特征提取,BERT(bidirectional encoder representations from transformers)用于文本特征提取。与许多现有多模态模型倾向使用的单流架构(即通过共享的 Transformer 联合编码视觉与文本输入)不同,本文刻意选择了双流架构。这种架构选择并非任意为之,而是源于多模态篡改检测与定位任务的独特语义需求。具体而言,篡改检测的核心在于识别跨模态的局部不一致性,例如图像某一细节区域与文本中某个词语之间的偏差。如果过早将两种模态强行投影到统一嵌入空间,往往会在全局对齐的过程中稀释掉模态内部的结构特征,导致篡改信号被淹没。相比之下,双流架构能够保持各模态的固有语义结构,在保证表征完整性的同时,为后续的跨模态对比提供更清晰的语义边界。更重要的是,双流架构与本文提出的局部到局部对比学习机制天然契合。该机制依赖于图像块与文本词元既相对独立又在任务驱动下实现语义对齐的表征方式,在处理涉及双模态局部篡改的场景时,这种分离再对齐的过程能够实现更细粒度、更精确的匹配,从而有效识别被篡改的区域或词元。换言之,双流架构不仅保留了模态内的真实性信号,还为跨模态篡改检测提供了必要的语义分辨率。图像编码器记做$E_i$,文本编码器记做$E_t$。输入图像$\boldsymbol{I}$被划分为$N$个$16\times16$像素的patch,并前置一个[*CLS*]标记;同理,输入文本$T$被划分为$M$个token,同样前置一个[*CLS*]标记。图像特征$E_i(\boldsymbol{I})$与文本特征$E_t(T)$可表示为

$$\begin{aligned}E_i(\boldsymbol{I})&=\{\boldsymbol{v}_{\text{cls}},\boldsymbol{v}_{\text{pat}}\}\\E_t(T)&=\{\boldsymbol{t}_{\text{cls}},\boldsymbol{t}_{\text{tok}}\}\end{aligned}\tag{1}$$

式中,$\boldsymbol{v}_{\text{cls}}$表示图像的[*CLS*]嵌入,$\boldsymbol{v}_{\text{pat}}=\{\boldsymbol{v}_{\text{pat}_1},\boldsymbol{v}_{\text{pat}_2},\cdots,\boldsymbol{v}_{\text{pat}_N}\}$对应$N$个patch的嵌入,$\boldsymbol{t}_{\text{cls}}$表示文本的[*CLS*]嵌入,$\boldsymbol{t}_{\text{tok}}=\{\boldsymbol{t}_{\text{tok}_1},\boldsymbol{t}_{\text{tok}_2},\cdots,\boldsymbol{t}_{\text{tok}_M}\}$对应$M$个token的嵌入。

### 1.3 篡改对比学习

近年来兴起的统一空间对齐方法(Radford 等,2021;Kim 等,2021)通过在大规模图文数据上进行预训练,将图像和文本映射到共享的嵌入空间,从而学习到通用的跨模态表示。这类方法在开放域检索和匹配任务中表现突出,但其对齐目标主要停留在全局语义一致性,缺乏对局部细粒度对齐和篡改痕迹建模的能力。在多模态篡改检测场景中,这一不足尤为突出。图像伪造往往集中于某些局部区域(如面部修改),而文本篡改通常涉及个别词汇或短语。在这种情况下,大部分区域在语义上依旧保持一致,如果强行依赖全局对齐来区分篡改与真实,不仅容易掩盖局部不一致性,还可能导致对比学习目标与篡改检测任务脱节。更进一步,即便图像与文本均被篡改,它们在整体语义上仍可能保持高度一致,从而进一步加剧了细粒度检测的难度。

为克服多模态对比学习中全局语义对齐的局限性,本文提出了一种面向篡改感知的局部—局部对比学习机制,实现了区别于统一嵌入方法的任务驱动隐式对齐。与以往基于像素到词或图块到词标记的统一匹配策略不同,本文方法引入了基于任务的、区域感知的监督机制,依托于篡改标注信息。对于每一个图文对,基于图像的边界框注释与文本的伪造位置,生成图像图块层面与文本词元层面的篡改掩码,从而使模型能够区分图像与文本中的篡改与非篡改区域,并在此基础上实现针对性的对比监督。随后,计算每一个图像图块与每一个文本词标记之间的两两相似度矩阵,并结合由篡改掩码引导的标签矩阵进行监督,从而显式编码细粒度的语义一致性与不一致性。与传统的统一嵌入空间不同,这种设计使模型不再仅追求模态间的整体匹配,而是聚焦于篡改敏感区域的跨模态对齐,从而在细粒度篡改场景下展现出更强的判别力。

如图 2(a)所示,首先基于图像与文本的全局特征标记$\boldsymbol{v}_{\text{cls}}$与$\boldsymbol{t}_{\text{cls}}$之间的相关性,计算对比学习目标函

数。其中，图像到文本的全局对全局篡改对比损失定义为

$$L_{\mathrm{v2t}}(\boldsymbol{v}_{\mathrm{cls}}, \boldsymbol{t}^{+}_{\mathrm{cls}}, \boldsymbol{t}^{-}_{\mathrm{cls}}) = \mathbb{E}_{p(I,T)}\left[ -\log \frac{\exp\left(Sim(\boldsymbol{v}_{\mathrm{cls}}, \boldsymbol{t}^{+}_{\mathrm{cls}})/\tau\right)}{\sum_{k=1}^{K}\exp\left(Sim(\boldsymbol{v}_{\mathrm{cls}}, \boldsymbol{t}^{-}_{\mathrm{cls}_k})/\tau\right)} \right] \tag{2}$$

式中，参数$\tau$为可学习的温度系数，用于调节对比学习分布的平滑性。$\boldsymbol{t}^{+}_{\mathrm{cls}}$表示与图像全局特征$\boldsymbol{v}_{\mathrm{cls}}$在语义上匹配的正样本文本特征，而$\boldsymbol{t}^{-}_{\mathrm{cls}}$表示与$\boldsymbol{v}_{\mathrm{cls}}$在语义上不一致的一组负样本文本特征。函数$Sim(\cdot)$通过内积运算计算图像与文本标记之间的相似度。类似地，文本到图像的对比损失表示为$L_{\mathrm{t2v}}(\boldsymbol{t}_{\mathrm{cls}}, \boldsymbol{v}^{+}_{\mathrm{cls}}, \boldsymbol{v}^{-}_{\mathrm{cls}})$，图像到图像的模态内对比损失表示为$L_{\mathrm{v2v}}(\boldsymbol{v}_{\mathrm{cls}}, \boldsymbol{v}^{+}_{\mathrm{cls}}, \boldsymbol{v}^{-}_{\mathrm{cls}})$，文本到文本的对比损失表示为$L_{\mathrm{t2t}}(\boldsymbol{t}_{\mathrm{cls}}, \boldsymbol{t}^{+}_{\mathrm{cls}}, \boldsymbol{t}^{-}_{\mathrm{cls}})$。综上，整体的“全局—全局”篡改对比损失函数可表示为

$$L_p(\boldsymbol{I},\boldsymbol{T}) = L_{\mathrm{v2t}}(\boldsymbol{v}_{\mathrm{cls}}, \boldsymbol{t}^{+}_{\mathrm{cls}}, \boldsymbol{t}^{-}_{\mathrm{cls}}) + L_{\mathrm{t2v}}(\boldsymbol{t}_{\mathrm{cls}}, \boldsymbol{v}^{+}_{\mathrm{cls}}, \boldsymbol{v}^{-}_{\mathrm{cls}}) + L_{\mathrm{v2v}}(\boldsymbol{v}_{\mathrm{cls}}, \boldsymbol{v}^{+}_{\mathrm{cls}}, \boldsymbol{v}^{-}_{\mathrm{cls}}) + L_{\mathrm{t2t}}(\boldsymbol{t}_{\mathrm{cls}}, \boldsymbol{t}^{+}_{\mathrm{cls}}, \boldsymbol{t}^{-}_{\mathrm{cls}}) \tag{3}$$

如图2(b)所示，在图文对中对图像块和文本标记的局部嵌入进行细粒度的篡改对比学习。首先根据图像中被篡改的边界框区域，将图像块的嵌入分类为正样本和负样本。对于每个图像块嵌入，如果该图像块位于篡改区域内，则标记为负样本；否则，标记为正样本。例如，图像块嵌入集合可以表示为$\{\boldsymbol{v}^{+}_{\mathrm{pat}_1}, \boldsymbol{v}^{-}_{\mathrm{pat}_2}, \boldsymbol{v}^{+}_{\mathrm{pat}_3}, \cdots, \boldsymbol{v}^{+}_{\mathrm{pat}_N}\}$，其中，$\boldsymbol{v}^{+}_{\mathrm{pat}_i}$表示位于非篡改区域的正样本，$\boldsymbol{v}^{-}_{\mathrm{pat}_i}$表示位于篡改区域的负样本。类似地，对于文本，根据篡改词汇的位置将标记嵌入分类为正样本和负样本。对于每个文本标记嵌入，如果对应的词汇已被修改，则标记为负样本；否则，标记为正样本。因此，文本标记嵌入集合可以表示为$\{\boldsymbol{t}^{+}_{\mathrm{tok}_1}, \boldsymbol{t}^{+}_{\mathrm{tok}_2}, \boldsymbol{t}^{-}_{\mathrm{tok}_3}, \cdots, \boldsymbol{t}^{+}_{\mathrm{tok}_M}\}$，其中，$\boldsymbol{t}^{+}_{\mathrm{tok}j}$表示未篡改的正样本，$\boldsymbol{t}^{-}_{\mathrm{tok}j}$表示篡改后的负样本。利用这些局部正负样本，定义图文对的“局部—局部”篡改对比损失为

$$L_q = L\left(\boldsymbol{v}^{+}_{\mathrm{pat}}, \boldsymbol{v}^{-}_{\mathrm{pat}}, \boldsymbol{t}^{+}_{\mathrm{tok}}, \boldsymbol{t}^{-}_{\mathrm{tok}}\right) = \mathbb{E}_{p(I,T)}\left[ -\log \frac{\exp\left(Sim\left(\boldsymbol{v}^{+}_{\mathrm{pat}}, \boldsymbol{t}^{+}_{\mathrm{tok}}\right)/\tau\right)}{A} \right]$$

$$A = \exp\left(Sim\left(\boldsymbol{v}^{+}_{\mathrm{pat}}, \boldsymbol{t}^{-}_{\mathrm{tok}}\right)/\tau\right) + \exp\left(Sim\left(\boldsymbol{v}^{-}_{\mathrm{pat}}, \boldsymbol{t}^{+}_{\mathrm{tok}}\right)/\tau\right) + \exp\left(Sim\left(\boldsymbol{v}^{-}_{\mathrm{pat}}, \boldsymbol{t}^{-}_{\mathrm{tok}}\right)/\tau\right) \tag{4}$$

式中，$\boldsymbol{v}^{+}_{\mathrm{pat}}$表示位于图像篡改区域外的正样本，而$\boldsymbol{v}^{-}_{\mathrm{pat}}$表示位于篡改区域内的负样本。类似地，$\boldsymbol{t}^{+}_{\mathrm{tok}}$表示未篡改的正文本样本，$\boldsymbol{t}^{-}_{\mathrm{tok}}$表示已篡改的负文本样本。该损失函数的目标是通过最小化正样本对之间的距离，并最大化负样本对之间的距离，从而提高图像块与文本标记之间的语义一致性。通过这种方式，模型能够有效地捕捉篡改在图像和文本局部尺度上的具体影响，从而提升篡改定位性能，特别是在涉及局部篡改的情况下。通过联合考虑“全局—全局”篡改对比损失和“局部—局部”篡改对比损失，最终的损失函数可以表示为

$$L_{\mathrm{mc}} = L_p + L_q \tag{5}$$

从理论视角来看，本文提出的“全局—全局”与“局部—局部”篡改对比学习机制可以分别视为语义一致性约束与局部扰动约束的组合。前者确保图像与文本在整体语义层面保持对齐；后者则通过显式利用篡改区域标注，使模型在细粒度尺度上区分真实与伪造内容。这样的双重约束不仅提升了特征表示的判别性，还在训练过程中引入了额外的正则化效应。对比学习中的InfoNCE(information noise-contrastive estimation)损失可以近似最大化视图间的互信息，从而提升表征的泛化能力。结合这一理论基础，本文的“局部—局部”对比损失项可理解为在互信息约束之外引入局部层级的扰动监督，有效降低模型对无关区域的依赖。换言之，最小化局部对比损失在本质上等价于减少模型的泛化误差界，使得训练过程具备更强的鲁棒性与泛化性。因此，该机制能够从理论层面解释本文方法在跨模态细粒

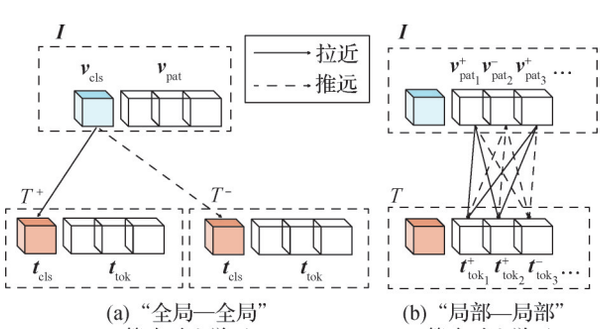



图2 “全局—全局”篡改对比学习与“局部—局部”篡改对比学习的比较

Fig. 2 Comparison between global-global manipulation contrastive learning and local-local manipulation contrastive learning((a)global-global manipulation contrastive learning; (b)local-local manipulation contrastive learning)

度篡改检测与定位任务中的优越表现。

### 1.4　多模态特征融合

#### 1.4.1　伪造感知交互模块

在提取图像特征 $\boldsymbol{E}_i(\boldsymbol{I})$ 和文本特征 $\boldsymbol{E}_t(\boldsymbol{T})$ 后，引入了双流跨模态注意力机制（Vaswani 等，2017），通过将其中一种模态作为查询（$\boldsymbol{Q}$），另一种模态作为键（$\boldsymbol{K}$）和值（$\boldsymbol{V}$），实现跨模态信息聚合，从而增强语义相关性。

在图像篡改定位任务中，将图像嵌入 $\boldsymbol{E}_i(\boldsymbol{I})$ 作为查询，文本嵌入 $\boldsymbol{E}_t(\boldsymbol{T})$ 则作为键和值。这种设计使得图像嵌入作为信息的主要来源，而文本嵌入提供补充细节，以增强图像特征的表达能力。该过程可以表示为

$$\boldsymbol{F}_{\mathrm{t2v}} = f_{\mathrm{cross\text{-}attn}}(\boldsymbol{E}_i(\boldsymbol{I}), \boldsymbol{E}_t(\boldsymbol{T}), \boldsymbol{E}_t(\boldsymbol{T})) \tag{6}$$

式中，集合 $\boldsymbol{F}_{\mathrm{t2v}} = \{i_{\mathrm{cls}}, i_{\mathrm{pat}}\}$，$i_{\mathrm{cls}}$ 表示 $[CLS]$ 嵌入，$i_{\mathrm{pat}} = \{i_{\mathrm{pat}_1}, i_{\mathrm{pat}_2}, \cdots, i_{\mathrm{pat}_N}\}$ 表示经过文本信息增强的 $N$ 个图像块嵌入。实验结果表明，通过补充文本嵌入实现的图像篡改定位提升效果是有限的。为了进一步提高定位精度，受 Wang 等人（2022）的启发，设计了伪造感知交互模块（forgery-aware interaction module，FAIM），如图 3 所示。

为了捕捉多个尺度上的篡改模式，引入了一个多尺度 Transformer，它在不同尺寸的图像块上进行操作。以先前图像编码的输出 $\boldsymbol{E}_i(\boldsymbol{I})$ 作为输入，将其分割成不同尺寸的空间图像块，并在不同的头部中计算图像块的自注意力。首先，从 $\boldsymbol{E}_i(\boldsymbol{I})$ 中重塑并提取形状为 $r_h \times r_h \times C$ 的图像块，再将其重塑为一维向量，以供第 $h$ 头使用。之后，使用全连接层将这些展平的向量嵌入为查询嵌入 $\boldsymbol{Q}_h \in \mathbf{R}^{N \times C_h}$，其中 $N = (H/4r_h) \times (W/4r_h)$，且 $\boldsymbol{C}_h = r_h \times r_h \times C$，类似的操作用于获取键嵌入 $\boldsymbol{K}_h$ 和值嵌入 $\boldsymbol{V}_h$。然后，计算注意力矩阵 $\boldsymbol{M}_i^h$，具体为

$$\boldsymbol{M}_i^h = f_{\mathrm{softmax}}\left(\frac{\boldsymbol{Q}_i^h (\boldsymbol{K}_i^h)^{\mathrm{T}}}{\boldsymbol{C}_h}\right)\boldsymbol{V}_i^h \tag{7}$$

然后，$\boldsymbol{M}_i^h$ 被上采样到原始图像的空间分辨率。最后，将所有头部的输出特征拼接起来，并输入到后续的 2D 残差块中进行处理，以获得输出 $\boldsymbol{W}_i \in \mathbf{R}^{(H/4)\times(W/4)\times C}$。已有研究表明，经过 JPEG（joint photographic experts group）压缩等压缩方法处理后，篡改图像的伪影变得不再显著，因此，在频率域中提取篡改特征，以补充 RGB 特征。首先，沿空间维度应用二维快速傅里叶变换（two-dimensional fast Fourier transform，2D FFT）将 $E_i(\boldsymbol{I})$ 转换到频率域并获得频谱表示 $\boldsymbol{F}(E_i(\boldsymbol{I})) \in \mathbf{R}^{H/4 \times W/4 \times C}$，接着，将 $\boldsymbol{F}(E_i(\boldsymbol{I}))$ 与一个可学习的滤波器 $\boldsymbol{G}_i \in \mathbf{R}^{H/4 \times W/4 \times C}$ 相乘，以建模不同频率带分量之间的依赖关系，具体为

$$\hat{\boldsymbol{G}}_i = \boldsymbol{G}_i \odot \boldsymbol{F}(E_i(\boldsymbol{I})) \tag{8}$$

式中，$\odot$ 表示哈达玛积。最后，执行逆快速傅里叶变换，将 $\hat{\boldsymbol{G}}_i$ 转换回空间域，得到频率感知特征 $\boldsymbol{X}_i$。给定 RGB 特征 $\boldsymbol{W}_i$ 和频率特征 $\boldsymbol{X}_i$，首先使用 1 ×1 卷积 $conv_q$、$conv_k$、$conv_v$，将 $\boldsymbol{W}_i$ 和 $\boldsymbol{X}_i$ 分别嵌入为查询（$\boldsymbol{Q}$）、键（$\boldsymbol{K}$）和值（$\boldsymbol{V}$）。然后，沿空间维度展平它们，得到 2D 嵌入 $\tilde{\boldsymbol{Q}}, \tilde{\boldsymbol{K}}, \tilde{\boldsymbol{V}} \in \mathbf{R}^{(HW/16)\times C}$，并计算融合后的特征，具体为

$$\tilde{\boldsymbol{Z}}_i = f_{\mathrm{softmax}}\left(\frac{\tilde{\boldsymbol{Q}}\tilde{\boldsymbol{K}}^{\mathrm{T}}}{\sqrt{H/4 \times W/4 \times C}}\right)\tilde{\boldsymbol{V}} \tag{9}$$

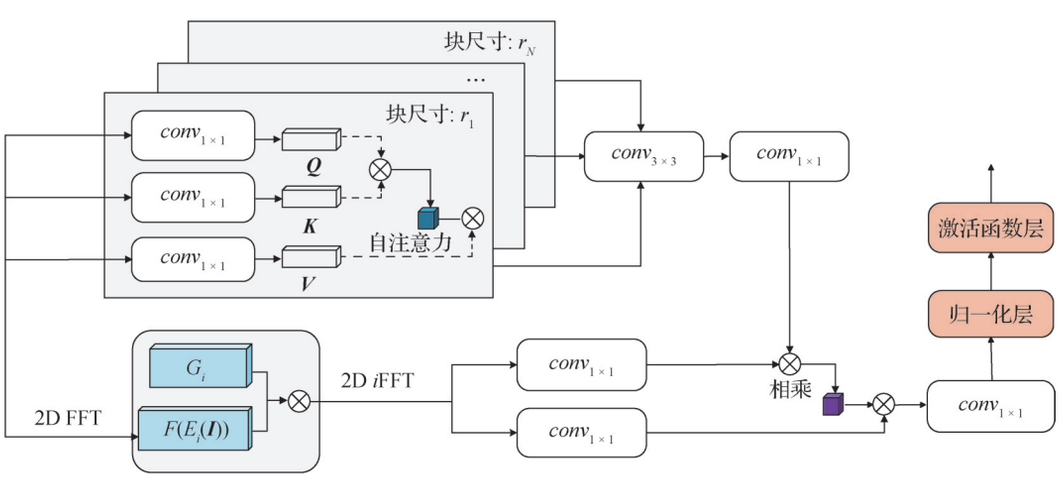



图 3　伪造感知交互模块

Fig. 3　Forgery-aware interaction module

式中，$\tilde{\boldsymbol{Z}}_i$结合了RGB域和频率域中的篡改特征，显著提升了模型在图像中定位篡改区域的能力。与后续的门控融合模块不同，伪造感知交互模块的目标在于在同一模态内部（图像）充分整合多源特征，尤其是通过跨域交叉注意力机制将RGB特征与频域特征进行对齐和互补，从而增强模型对图像细粒度篡改痕迹的捕捉能力。该过程属于同源模态的伪造增强交互，强调的是对单一视觉模态的深度建模和伪造感知。

图像的局部嵌入捕捉了细粒度的空间信息，这对于涉及图像篡改定位的任务尤为重要。受到HAMMER中提出的局部块注意力聚合（LPAA）策略（Shao等，2023）的启发，从局部图像嵌入中通过注意力计算来聚合特定于篡改的特征，以提高定位精度。局部图像嵌入的聚合过程可表示为

$$(\boldsymbol{u}_{\text{cls}}, \boldsymbol{u}_{\text{pat}}) = \boldsymbol{F}_{\text{t2v}} + \tilde{\boldsymbol{Z}}_i \tag{10}$$

$$\boldsymbol{l}_{\text{cls}} = f_{\text{cross-attn}}([agg], \boldsymbol{u}_{\text{pat}}, \boldsymbol{u}_{\text{pat}}) \tag{11}$$

式中，[*agg*]为聚合标记，用于聚合局部图像块的空间信息。

1.4.2　跨模态门控融合模块

对于文本篡改定位，构建了一个多模态聚合器。首先通过自注意力层对文本嵌入进行优化，得到的表示被用于查询，而图像嵌入$\boldsymbol{E}_i(\boldsymbol{I})$提供键和值表示，从而实现文本和视觉线索的有效融合。该过程表示为

$$\boldsymbol{F}_{\text{v2t}} = f_{\text{cross-attn}}\left(f_{\text{self-attn}}\left(\boldsymbol{E}_t(T)\right), \boldsymbol{E}_i(\boldsymbol{I}), \boldsymbol{E}_i(\boldsymbol{I})\right) \tag{12}$$

式中，$\boldsymbol{F}_{\text{v2t}} = \{\boldsymbol{m}_{\text{cls}}, \boldsymbol{m}_{\text{tok}}\}$，其中$\boldsymbol{m}_{\text{cls}}$是融合后的[*CLS*]标记，$\boldsymbol{m}_{\text{tok}} = \{\boldsymbol{m}_{\text{tok}_1}, \boldsymbol{m}_{\text{tok}_2}, \cdots, \boldsymbol{m}_{\text{tok}_M}\}$表示经过视觉信息增强的$M$个标记嵌入序列。

仅依赖多模态聚合器产生的融合的[*CLS*]标记进行二分类和多标签分类是次优的。这个局限性源于[*CLS*]标记主要是以语言为中心的，将视觉流视为辅助信息，而这两种分类任务都需要显式且具区分性的视觉线索。为了弥补这一缺陷，设计了一个跨模态门控融合模块（CGFM），如图4所示，该模块学习一个动态门控向量来调节每种模态的贡献，从而生成任务感知的多模态表示。与前面的伪造感知交互模块不同，门控融合模块的重点在于在异构模态之间（图像与文本）实现动态特征融合。通过引入sigmoid函数学习的门控因子，模型能够为不同模态分配自适应的权重，从而在跨模态层面上缓解语义不一致带来的干扰，提升多任务检测的鲁棒性。该过程属于跨源模态的动态加权融合，强调的是跨模态信息交互与统一表征。

经观察发现，检测头部中的跨注意力权重反映了不同图像篡改定位图像块的相对重要性以及它们与篡改图像块的相关性。具体而言，对局部块注意力聚合部分的跨注意力的权重$W_c$做softmax函数，以将权重值限制在0到1之间。该权重表示由检测器识别的篡改区域的可能性。具体为

$$G = f_{\text{softmax}}(W_c) \tag{13}$$

跨模态门控融合模块视觉部分的最终输入是局部块注意力聚合的输出与权重$G$的点积。即

$$\boldsymbol{f}_{\text{cls}} = G \cdot \boldsymbol{l}_{\text{cls}} \tag{14}$$

随后，将$\boldsymbol{f}_{\text{cls}}$与多模态聚合器的[*CLS*]标记$\boldsymbol{m}_{\text{cls}}$进行跨模态门控融合，得到融合的表达式为

$$S_{\text{gate}} = f_{\text{sigmoid}}(W_l \boldsymbol{f}_{\text{cls}} + W_v \boldsymbol{m}_{\text{cls}}) \tag{15}$$

$$\boldsymbol{g}_{\text{cls}} = S_{\text{gate}} \boldsymbol{f}_{\text{cls}} + (1 - S_{\text{gate}}) \boldsymbol{m}_{\text{cls}} \tag{16}$$

式中，$W_l$和$W_v$是可学习的参数。$\boldsymbol{g}_{\text{cls}}$有效地整合了视觉和语言模态的信息，随后输入到多层感知机（MLP）中，用于多模态检测任务。

### 1.5　多任务学习

本文框架在以下4个方面具有多功能性，能够检测和定位多模态虚假信息：即图像深度伪造定位、

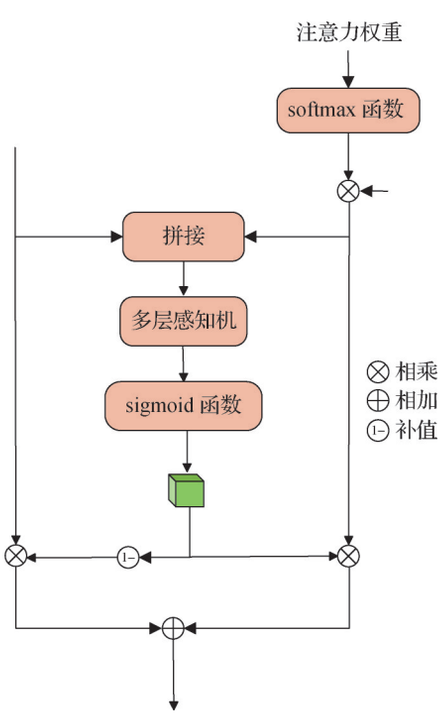



图4　跨模态门控融合模块

Fig. 4　Cross-modal gated fusion module

文本篡改定位、多模态深度伪造检测以及深度伪造细粒度分类。这些目标通过不同的训练监督来实现。

1.5.1　图像深度伪造定位损失

为了定位已被篡改的图像区域，设计了一个边界框检测头，该检测头由3层 MLP 组成。将边界框检测头表示为$D_v$，它以局部块注意力聚合的输出$l_{\text{cls}}$作为输入，并进行相应的边界框预测，该过程由定位目标进行监督，具体为

$$L_{\text{ig}} = \mathbb{E}_{(I,T)\sim p}\Big[\Big\| f_{\text{sigmoid}}(D_v(\boldsymbol{l}_{\text{cls}})) - y_{\text{box}} \Big\| + L_{\text{GIoU}}(f_{\text{sigmoid}}(D_v(\boldsymbol{l}_{\text{cls}})) - y_{\text{box}})\Big] \tag{17}$$

式中，$y_{\text{box}}$是边界框检测的真实标签。

1.5.2　文本篡改定位损失

为了定位篡改的文本标记，将文本表示$m_{\text{tok}}$输入到一个文本篡改检测器中，用于检测篡改位置。与现有工作类似，还采用了聚合器和检测器的动量版本，记做$\hat{D}_t$，整体目标函数计算为

$$L_{\text{tg}} = \mathbb{E}_{(I,T)\sim p}[(1-\gamma)H\big(D_t(m_{\text{tok}}), y_{\text{tok}}\big) + \gamma KL(D_t(m_{\text{tok}}) \Big\| \hat{D}_t(\hat{m}_{\text{tok}}))\,] \tag{18}$$

式中，$H(\cdot)$表示交叉熵损失函数，$D_t$是由3层 MLP 组成的标记检测器，$\hat{m}_{\text{tok}}$是由动量更新的多模态聚合器提取的文本嵌入。$\gamma$是权重系数。

1.5.3　多模态深度伪造检测二分类损失

一旦通过图像与文本嵌入融合获得高层表征，即可用于推断给定视觉—语言对中是否存在跨模态的深度伪造内容。为实现这一目标，以跨模态门控融合模块输出$g_{\text{cls}}$标记作为输入，将其馈送至定制化的检测网络以执行二分类深度伪造推理。该检测网络设计为轻量级多层感知机（MLP），包含4个全连接层（fully connected layer，FC），其输出经 softmax 归一化后表征深度伪造概率。网络训练采用经典二分类损失函数进行监督，损失函数为

$$L_{\text{bc}} = \mathbb{E}_{(I,T)\sim p}H\Big(C_b\big(\boldsymbol{g}_{\text{cls}}\big), y_{\text{brl}}\Big) \tag{19}$$

式中，$y_{\text{brl}}$是图像—文本对的真实标签。

1.5.4　深度伪造细粒度分类损失

除了上述二分类深度伪造检测外，模型还能够进行细粒度的深度伪造鉴定，旨在确定具体的篡改类型，如面部/文本交换（face swap/text swap，FS/TS）或面部/文本属性（face attribute/text attribute，FA/TA）篡改。这本质上是一个四分类任务，通过基于 MLP 的分类器实现，其学习目标函数为

$$L_{\text{mlc}} = \mathbb{E}_{(I,T)\sim p}H\Big(C_m\big(\boldsymbol{g}_{\text{cls}}\big), y_{\text{mrl}}\Big) \tag{20}$$

式中，$C_m$由4层 MLP 组成，$y_{\text{mrl}}$是多标签分类任务的真实标签。

通过结合所提出的篡改对比学习损失$L_{\text{mc}}$和多任务学习目标，最终的损失函数为

$$L_{\text{total}} = \alpha_1 L_{\text{mc}} + \alpha_2 L_{\text{ig}} + \alpha_3 L_{\text{bc}} + \alpha_4 L_{\text{mlc}} + \alpha_5 L_{\text{tg}} \tag{21}$$

## 2　实　验

### 2.1　实验设置

所有实验均在一台 NVIDIA RTX 4090 GPU 上使用 PyTorch 框架（Paszke 等，2017）进行。输入图像被调整为256 × 256像素大小，文本序列的最大长度设置为50。图像和文本的嵌入分别通过 ViT 和 BERT 生成。训练进行了50个周期，批量大小为32。采用了 AdamW 优化器（Loshchilov 和 Hutter，2019），权重衰减设置为0.02。学习率在前5个周期内逐渐上升，达到$2 \times 10^{-5}$，然后根据 cosine_in_step 调度逐渐衰减至$1 \times 10^{-7}$。式（18）中的权重系数$\gamma$设置为0.4。式（21）中的损失函数系数配置为$\alpha_1 = \alpha_2 = 0.1$、$\alpha_3 = \alpha_4 = \alpha_5 = 1$。

### 2.2　数据集

实验分析基于3个数据集进行。1）DGM$^4$数据集（Shao 等，2023）是首个专门针对多模态篡改检测与定位的权威基准，提供了图像篡改区域和文本修改标记的细粒度标注，并涵盖图像篡改、文本篡改及二者混合等多种类型。该数据集规模超过20万对图文样本，具有较高的多样性和挑战性。2）Twitter（Boididou 等，2015）训练集包含4 992条真实推文和9 470条谣言推文，测试集包含1 215条真实推文和717条谣言推文，主要用于多模态虚假信息检测任务。3）Weibo（Jin 等，2017）训练集包含3 783条真实微博和3 749条谣言微博，测试集包含 996条真实微博和1 000条谣言微博，同样聚焦于跨模态虚假信息检测。

### 2.3　评价指标

在多模态多任务学习中，使用12个评估指标全面评估每个任务的表现。对于验证真实性的二分类任务，评估指标包括准确率（accuracy，ACC）、接收器操作特征曲线下的面积（area under the curve，AUC）和等错误率（equal error rate，EER）；对于篡改类型检

测的多标签分类任务，采用平均精度均值（mean average precision，mAP）、类别平均F1分数（class-wise F1 score，CF1）和总体F1分数（overall F1 score，OF1）作为评估指标。对于篡改图像的定位，评估标准包括平均交并比（mean intersection over union，IoUmean）以及在0.5和0.75阈值下的交并比IoU50和IoU75；对于篡改文本的定位，评估指标包括精确率（precision）、召回率（recall）和F1分数。在所有指标中，除了EER是较低的值表示较优的性能，其余指标均是较高的值表示系统性能较好。

### 2.4 实验结果与分析

#### 2.4.1 与多模态篡改检测模型的比较

为了评估所提方法的有效性，与4个最先进的多模态模型进行比较，包括ViLT（vision-language Transformer）（Kim 等，2021）、CLIP（contrastive language-image pre-training）（Radford等，2021）、VLP-GF（visual-language pre-training with gate fusion）（Zhang等，2024）和HAMMER（Shao等，2023）。

在DGM$^4$基准测试上的全面多任务结果如表1所示。可以看出，模型在多模态深度伪造检测二分类任务中的AUC为93.23%，在深度伪造细粒度多标签分类任务中的OF1为80.08%，图像深度伪造定位中的IoUmean为77.04%，在文本篡改定位中的F1为72.15%。本文方法在多模态篡改检测的4个子任务中均取得了优于现有方法的性能，尤其在细粒度检测与定位方面优势显著。与统一空间对齐的ViLT和CLIP相比，在所有指标上均有明显提升，表明单纯依赖全局一致性不足以应对篡改检测，而引入双流结构与“局部—局部”对比学习能够更有效地捕捉细粒度不一致性。在采用分层推理结构的HAMMER和引入门控机制的VLP-GF对比中，本文方法在图像定位和文本定位上表现突出，说明伪造感知交互模块（FAIM）在提升视觉定位精度上发挥了关键作用，而跨模态门控融合模块（CGFM）有效缓解了跨模态任务间的干扰，从而提升了文本篡改识别的召回率。

**表1 在DGM$^4$数据集上与先进方法的比较**

**Table 1 Comparison with state-of-the-art approaches on the DGM$^4$ dataset**

/%

| 方法 | 二分类 | | | 多标签分类 | | | 图像定位 | | | 文本定位 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | ACC | EER ↓ | mAP | OF1 | CF1 | IoUmean | IoU50 | IoU75 | precision | recall | F1 |
| CLIP | 83.22 | 76.40 | 24.61 | 66.00 | 62.31 | 59.52 | 49.51 | 50.03 | 38.79 | 58.12 | 22.11 | 32.03 |
| ViLT | 85.16 | 78.38 | 22.88 | 72.37 | 66.00 | 66.14 | 59.32 | 65.18 | 48.10 | 66.48 | 49.88 | 57.00 |
| VLP-GF | 92.84 | 86.13 | 14.45 | 85.65 | 79.07 | **80.02** | 76.73 | **83.89** | 76.24 | **76.42** | 66.80 | 71.29 |
| HAMMER* | 92.70 | 85.39 | 15.36 | 85.29 | 78.80 | 77.93 | 73.35 | 81.33 | 71.44 | 74.83 | 66.00 | 70.14 |
| 本文 | **93.23** | **86.29** | **14.22** | **85.97** | **80.08** | 79.00 | **77.04** | 83.03 | **77.85** | 72.66 | **71.65** | **72.15** |

注：加粗字体表示各列最优结果。*表示在本文实验环境下复现的结果，↓表示值越小越好。

进一步地，如图5所示，模型在深度伪造细粒度多标签分类任务中对每种篡改类型的F1分数均高于HAMMER，证明了模型整体检测性能的稳健性，突出了其在细粒度检测中的优势。部分成功检测和定位的可视化示例如图6所示，其中，红色区域表示真实标签，蓝色区域表示模型预测结果。

#### 2.4.2 与单模态篡改检测模型的对比分析

为进一步系统性评估模型性能，分别在两种单模态伪造数据划分中，将所提方法与两组单模态基线模型进行对比。为确保公平性，在基线模型中增加了定位头模块，使其能够同步执行分类与定位任

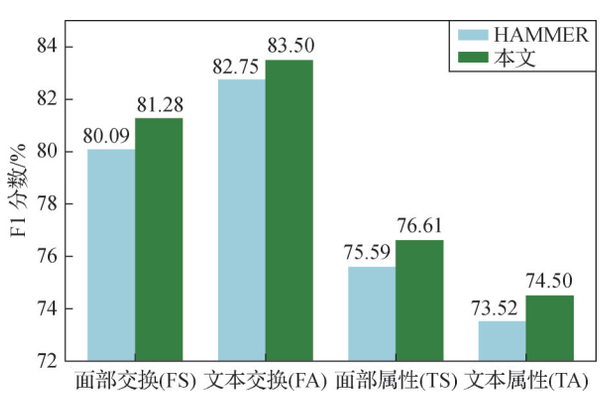



图5 4种细粒度操作类型的分类F1分数

Fig. 5 The F1 scores for the classification of four fine-grained manipulation types

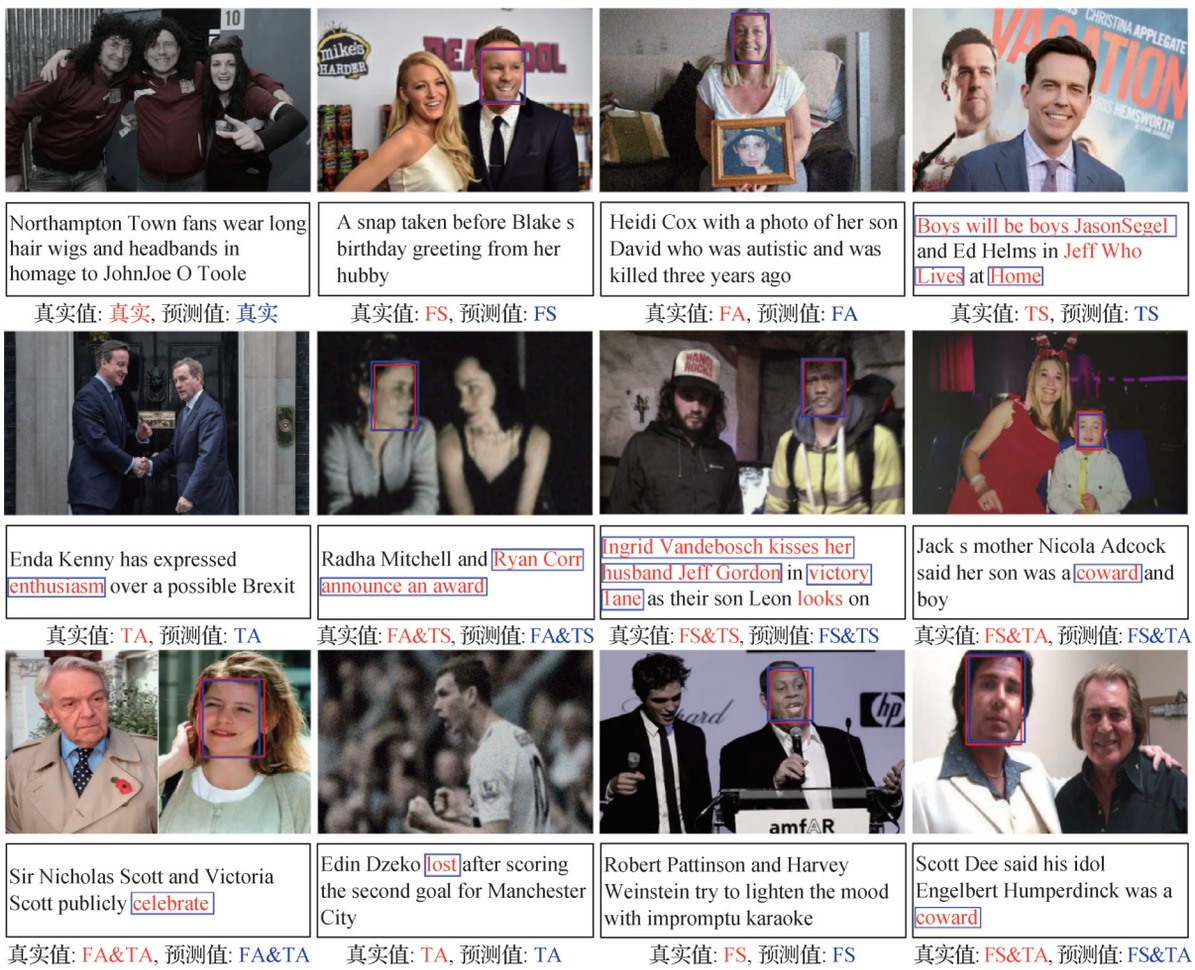



图 6　DGM⁴数据集中部分样本的可视化结果

Fig. 6　Visualization of some samples on the DGM⁴ dataset

务。在纯图像篡改检测方面，与 TS（two-stream）（Luo 等，2021）、MAT（multi-attentional deepfake detection）（Zhao 等，2021）、VLP-GF（Zhang 等，2024）及 ViKI（vision-language knowledge interaction）（Li 等，2024）进行对比；在纯文本篡改检测方面，选取 BERT（Devlin 等，2019）、LUKE（language understanding with knowledge-based embeddings）（Yamada 等，2020）、VLP-GF（Zhang 等，2024）和 ViKI（Li 等，2024）作为对比基准。需要说明的是，VLP-GF 与 ViKI 原为多模态模型，本研究将其适配至单模态场景。结果如表 2 和表 3 所示。可以看出，在单模态伪造检测任务中，本文方法的性能显著优于现有单模态基线。这一优势得益于框架在保持模态特异性表征的同时，能够通过跨模态交互机制挖掘互补线索，从而增强模型对细粒度篡改模式的建模能力。实验结果显示，该方法不仅在二分类准确率上取得了提升，更在篡改区域的定位精度上表现突出。这说明，通过多模态媒体数据训练而得的表征在单一模态任务中同样具备优势，能够更有效地捕捉局部篡改痕迹。这些结果不仅验证了多模态学习在传统单模态场景中的潜在价值，也进一步证明了本文提出框架的广泛

**表 2　DGM⁴数据集上图像深度伪造检测方法的比较**

**Table 2　Comparison of image deepfake detection methods on the DGM⁴ dataset**

/%

| 方法 | 二分类 | | | 图像定位 | | |
|---|---|---|---|---|---|---|
| | AUC | ACC | EER ↓ | IoUmean | IoU50 | IoU75 |
| TS | 91.80 | 82.89 | 17.11 | 72.85 | 79.12 | 74.06 |
| MAT | 91.31 | 82.36 | 17.65 | 72.88 | 78.98 | 74.70 |
| VLP-GF | 91.61 | 84.64 | 15.95 | 72.85 | 80.81 | 67.88 |
| ViKI | 91.85 | 84.90 | 15.92 | 75.93 | 82.16 | 74.57 |
| 本文 | **92.90** | **85.33** | **14.25** | **76.10** | **82.56** | **78.03** |

注：加粗字体表示各列最优结果，↓表示值越小越好。

表3 DGM⁴数据集上序列标注方法的比较

Table 3 Comparison of sequence tagging methods on the DGM⁴ dataset

/%

| 方法 | 二分类 | | | 文本定位 | | |
|---|---|---|---|---|---|---|
| | AUC | ACC | EER ↓ | precision | recall | F1 |
| BERT | 80.82 | 68.98 | 28.02 | 41.39 | 63.85 | 50.23 |
| LUKE | 81.39 | 76.81 | 27.88 | 50.52 | 37.93 | 43.33 |
| VLP-GF | 91.66 | 84.47 | 16.20 | 72.91 | 64.50 | 68.45 |
| ViKI | 92.31 | 85.35 | 15.27 | **78.46** | 65.09 | **71.15** |
| 本文 | **93.38** | **87.00** | **14.33** | 69.82 | **71.63** | 70.71 |

注:加粗字体表示各列最优结果,↓表示值越小越好。

表4 在跨数据集上与先进方法的比较

Table 4 Comparison with state-of-the-art methods on the cross-dataset

/%

| 方法 | Twitter数据集 | | | Weibo数据集 | | |
|---|---|---|---|---|---|---|
| | AUC | ACC | EER ↓ | AUC | ACC | EER ↓ |
| HAMMER | 64.83 | 55.30 | 41.43 | 56.96 | 50.32 | 43.71 |
| 本文 | **76.25** | **65.32** | **28.22** | **60.86** | **56.65** | **41.02** |

注:加粗字体表示各列最优结果,↓表示值越小越好。

适用性与强泛化能力。

2.4.3 在跨数据集上与先进方法的对比

本节对比分析了HAMMER与本文方法在跨数据集场景下的泛化表现。选择在两个典型的多模态谣言检测数据集Twitter与Weibo上开展跨数据集实验,结果如表4所示。本文方法在二分类指标上均优于HAMMER,这一结果充分验证了所提模型的跨数据集泛化能力。

### 2.5 消融实验

2.5.1 关键模块的消融研究

为系统评估各模块的独立贡献与整体协同效果,在实验设置下进行了关键模块的消融实验,结果如表5所示。可以发现:首先,引入“局部—局部”的篡改对比学习机制显著提升了模型在多任务场景下的整体表现。这说明仅依赖全局语义对齐难以充分捕捉细粒度篡改痕迹,而该机制能够显式建模图像块与文本词元之间的语义错配,有效强化了跨模态特征的辨别力。其次,伪造感知交互模块通过多层次聚合视觉特征,增强了模型对不同尺度伪造痕迹的敏感性。这一机制不仅能捕捉微小的人脸或区域篡改,还能兼顾大范围语义改动,从而在图像定位任务中显著提升精度。再次,跨模态门控融合模块在特征聚合时引入动态门控机制,使模型能够根据任务需求灵活分配模态权重。结果表明,该模块在真实性二分类与篡改类型多分类任务上均带来明显增益,验证了动态融合策略对提升模型判别力的重要作用。总体而言,所有模块的联合使用在各项任务中均取得了最优性能,不仅展示了各模块的独立有效性,更凸显了它们在整体架构中的互补性与必要性。这一结果充分验证了本文提出的多视角视觉—语言信息交互框架在应对复杂篡改检测任务中的优势与合理性。

表5 针对本文模型中关键模块的消融实验研究

Table 5 Ablation studies on the critical modules in the proposed model

/%

| 方法 | 二分类 | | | 多标签分类 | | | 图像定位 | | | 文本定位 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | ACC | EER ↓ | mAP | OF1 | CF1 | IoUmean | IoU50 | IoU75 | precision | recall | F1 |
| + $L_p$ | 92.93 | 85.83 | 14.61 | 85.88 | 79.54 | 78.57 | 74.58 | 82.54 | 73.37 | 72.30 | 71.15 | 71.69 |
| + $L_p$ + FAIM | 92.98 | 85.72 | 14.74 | 85.61 | 79.88 | 78.83 | **78.19** | **84.98** | **79.65** | 71.27 | **72.05** | 71.03 |
| + $L_p$ + CGFM | 93.05 | 86.03 | **14.22** | **86.27** | **80.17** | **79.19** | 75.77 | 82.68 | 76.77 | 72.38 | 71.16 | **72.15** |
| + $L_p$ + FAIM + CGFM | **93.23** | **86.29** | **14.22** | 85.97 | 80.08 | 79.00 | 77.04 | 83.03 | 77.85 | **72.66** | 71.65 | **72.15** |

注:加粗字体表示各列最优结果,↓表示值越小越好。

2.5.2 各模态作用的消融研究

为评估不同模态在多模态篡改检测与定位任务中的独立贡献,本研究通过选择性移除图像或文本模态进行消融实验,旨在探究辅助性跨模态信息缺失对模型整体性能的影响。两种单模态变体的架构设计如图7所示,其与完整多模态模型的性能对比数

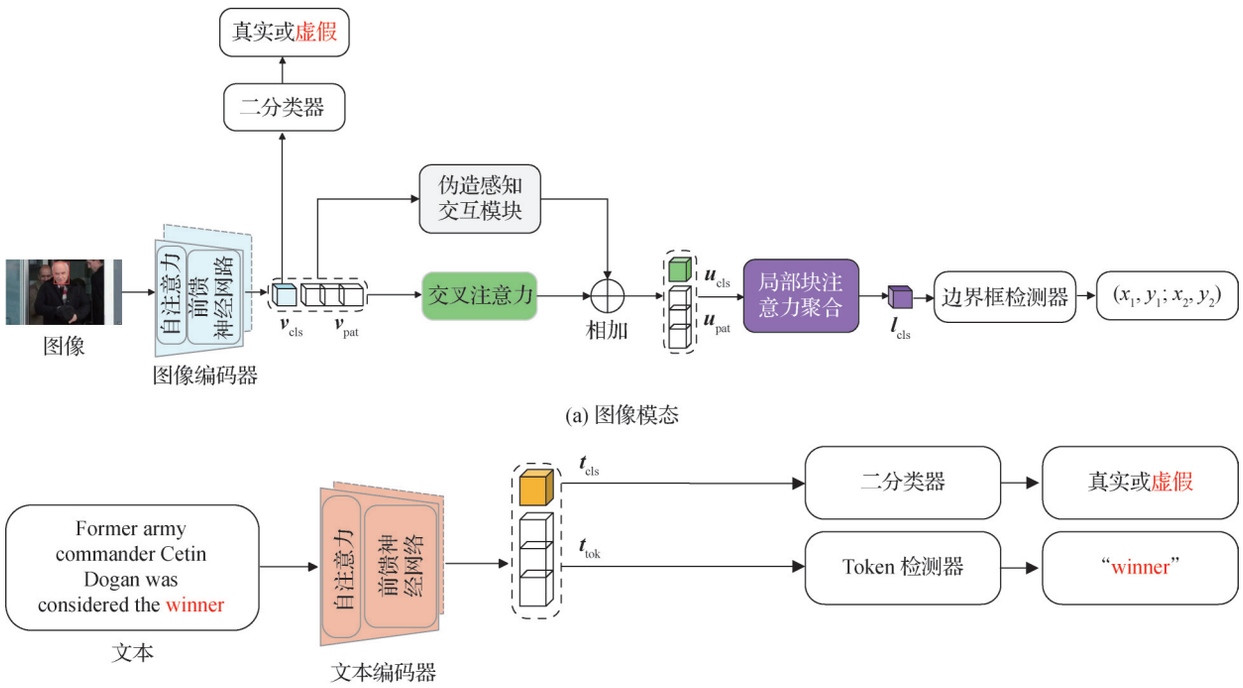



图 7　图像模态与文本模态的消融实验框架

Fig. 7　The ablation study framework for the image modality and the text modality ((a) image modality; (b) text modality)

据详见表 6 和表 7。实验结果表明，任一模态的缺失均会导致性能显著下降。视觉输入的移除会大幅降低文本篡改检测与定位的准确率，证实图像模态提供了关键的互补信息。这种跨模态交互机制使模型能够更有效地识别并定位篡改内容。反之，仅依赖文本输入则难以在篡改检测与定位任务中取得理想效果。

2.5.3　关于篡改对比学习的消融研究

为了分析“全局—全局”与“局部—局部”篡改对比学习损失的独立作用，设计了有针对性的消融实验，通过选择性地移除其中任一组件来评估其对整体性能的影响。结果如表 8 所示，表明无论去除哪一项损失都会导致多项任务中性能的下降。“全局—

**表 6　图像模态的消融研究结果**

**Table 6　Ablation study results of the image modality**

/%

| 方法 | 二分类 | | | 图像定位 | | |
|---|---|---|---|---|---|---|
| | AUC | ACC | EER ↓ | IoUmean | IoU50 | IoU75 |
| 图像模态 | 92.85 | 85.04 | 15.00 | 74.72 | 81.85 | 75.23 |
| 本文 | **92.90** | **85.33** | **14.25** | **76.10** | **82.56** | **78.03** |

注：加粗字体表示各列最优结果，↓表示值越小越好。

**表 7　文本模态的消融研究结果**

**Table 7　Ablation study results of the text modality**

/%

| 方法 | 二分类 | | | 文本定位 | | |
|---|---|---|---|---|---|---|
| | AUC | ACC | EER ↓ | precision | recall | F1 |
| 文本模态 | 63.88 | 62.43 | 45.54 | 43.45 | 35.78 | 36.42 |
| 本文 | **93.38** | **87.00** | **14.33** | **69.82** | **71.63** | **70.71** |

注：加粗字体表示各列最优结果，↓表示值越小越好。

全局”对比损失有效强化了跨模态的高层次语义对齐，使模型能够从整体语义层面区分真实与篡改样本；而“局部—局部”对比损失则更专注于捕捉输入中局部区域的细粒度不一致性，从而显著提升模型对篡改痕迹的敏感性与定位能力。将两者结合不仅实现了全局一致性与局部差异性的互补建模，还在多模态篡改检测与定位任务中带来了更鲁棒和精确的性能表现。

**表8　所提方法中篡改对比损失的消融研究结果**

**Table 8　Ablation study results of the manipulative contrastive loss in the proposed method**

/%

| 方法 | 二分类 | | | 多标签分类 | | | 图像定位 | | | 文本定位 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | ACC | EER↓ | mAP | OF1 | CF1 | IoUmean | IoU50 | IoU75 | precision | recall | F1 |
| 本文 w/o $L_g$ | 93.08 | 86.09 | 14.32 | 85.84 | 79.99 | 79.11 | 75.58 | 82.44 | 76.44 | **73.22** | 70.32 | 71.73 |
| 本文 w/o $L_p$ | 92.87 | **86.45** | **13.94** | 85.80 | 79.39 | **79.34** | 76.40 | 82.43 | 77.25 | 71.87 | 69.45 | 71.81 |
| 本文 | **93.23** | 86.29 | 14.22 | **85.97** | **80.08** | 79.00 | **77.04** | **83.03** | **77.85** | 72.66 | **71.65** | **72.15** |

注：加粗字体表示各列最优结果。w/o表示消去该组件，↓表示值越小越好。

### 2.5.4　伪造感知交互模块组件的消融研究

伪造感知交互模块由多尺度Transformer（multi- scale Transformer，MT）与频率滤波器（frequency filter，FF）两部分组成，分别用于增强模型对不同类型伪造信号的建模能力。MT通过在不同尺度的图像块之间建立全局依赖关系，能够有效捕捉多粒度的局部不一致性特征，从而提升模型在复杂伪造场景下的鲁棒性与定位精度；而FF则聚焦于频域信息建模，能够识别空间域难以显式检测的细微伪造痕迹。为验证两者的有效性，分别从伪造感知交互模块中移除MT与FF，并在DGM$^4$数据集上验证性能下降情况。定量结果如表9所示，移除任一组件都会导致性能下降，这表明，MT与FF在功能上具有高度互补性：前者加强了模型对局部空间伪造的敏感性，后者提供了频域层面的判别线索。二者协同作用使模型能够在多模态伪造检测任务中实现更全面、更稳健的性能提升。

### 2.5.5　伪造感知交互模块的有效性

为评估伪造感知交互模块的影响，对原始的局部块注意力聚合（LPAA）模型与引入伪造感知交互模块后的增强版本进行了对比实验，如图8所示。集成伪造感知交互模块的LPAA模型在所有IoU指标上均显著优于原始模型。该结果验证了伪造感知交互模块在提升模型空间特征表达能力方面的有效性，并显著增强了伪造区域的定位精度。

### 2.5.6　跨模态门控融合模块的有效性

为了验证跨模态门控融合模块的有效性，与常见的线性融合策略进行了对比实验，两种策略的结构如图9所示。

实验结果如表10所示。更进一步，在Twitter与Weibo数据集上对跨模态门控融合模块的效果进行了额外的消融实验，结果如表11所示。可以观察到，线性融合仅在固定权重下简单拼接或加权不同模态特征，难以充分建模任务依赖的模态贡献差异；

**表9　伪造感知交互模块组件的消融研究**

**Table 9　Ablation study on components of the forgery-aware interaction module**

/%

| 方法 | 二分类 | | | 多标签分类 | | | 图像定位 | | | 文本定位 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | ACC | EER↓ | mAP | OF1 | CF1 | IoUmean | IoU50 | IoU75 | precision | recall | F1 |
| w/o MT | 93.08 | 85.97 | 14.28 | 85.79 | 80.03 | **79.07** | 76.24 | 82.45 | 77.10 | 71.42 | **72.05** | 71.73 |
| w/o FF | 92.76 | 85.80 | 14.50 | 85.79 | 79.81 | 78.79 | 76.77 | 82.96 | 77.14 | **72.88** | 71.10 | 71.98 |
| 本文 | **93.23** | **86.29** | **14.22** | **85.97** | **80.08** | 79.00 | **77.04** | **83.03** | **77.85** | 72.66 | 71.65 | **72.15** |

注：加粗字体表示各列最优结果。w/o表示消去该组件，↓表示值越小越好。

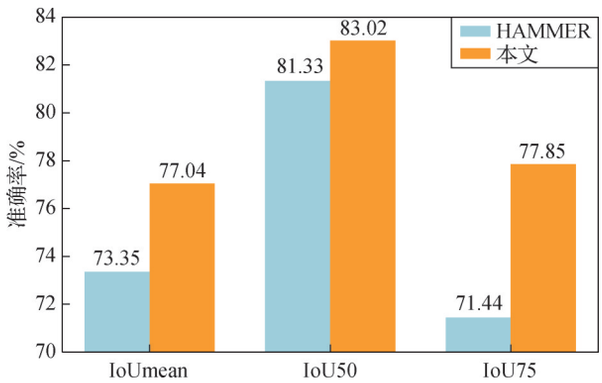
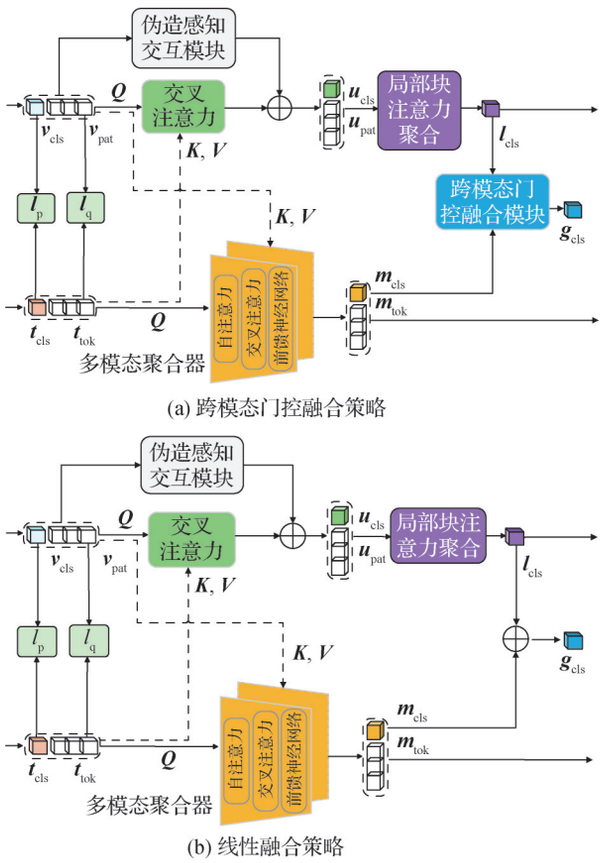



图 8　伪造感知交互模块的有效性

Fig. 8　Effectiveness of the forgery-aware interaction module




(a) 跨模态门控融合策略

(b) 线性融合策略

图 9　跨模态门控融合策略与线性融合策略结构对比

Fig. 9　Architectural comparison between cross-modal gated fusion strategy and linear fusion strategy ((a) cross-modal gated fusion strategy; (b) linear fusion strategy)

相比之下，跨模态门控融合模块通过引入动态门控机制，能够根据任务需求自适应地调节视觉与文本模态的贡献比例，从而实现更具判别力和更具语义感知的多模态嵌入表示。结果表明，采用门控融合策略后，无论是在二分类真实性检测还是多标签篡改类型识别任务中，模型的性能均显著优于线性融合。这一发现凸显了模型在复杂篡改场景下的必要性：其动态门控能力不仅避免了模态冗余信息的干扰，还有效强化了跨模态间的互补性，确保了模型能够生成面向任务优化的联合特征表示。

### 2.5.7　性能评估

为了验证所提模型的有效性，与其他 2 种先进方法进行比较，结果如表 12 所示。实验显示，虽然本文模型的参数规模相对更大，但 AUC 指标取得了最佳成绩。说明更高的参数容量提升了模型对复杂特征的表达与建模能力。此外，从浮点运算次数（floating point operations，FLOPs）的对比来看，模型在保证计算效率的前提下，能够充分发挥参数优势，从而带来性能的提升。

**表 10　跨模态门控融合策略与线性融合策略结果对比**

**Table 10　Comparative results of cross-modal gated fusion versus linear fusion strategies**

/%

| 方法 | 二分类 | | | 多标签分类 | | | 图像定位 | | | 文本定位 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | ACC | EER ↓ | mAP | OF1 | CF1 | IoUmean | IoU50 | IoU75 | precision | recall | F1 |
| 线性融合策略 | 92.97 | 85.72 | 14.73 | 85.61 | 79.88 | 78.82 | **78.15** | **84.92** | **79.25** | 71.22 | 71.58 | 71.13 |
| 本文 | **93.23** | **86.29** | **14.22** | **85.97** | **80.08** | **79.00** | 77.04 | 83.03 | 77.85 | **72.66** | **71.65** | **72.15** |

注：加粗字体表示各列最优结果，↓表示值越小越好。

## 3　结　论

本研究提出了一种创新的多模态篡改检测与定位模型。该模型通过多层级分析框架与对比学习策略的协同优化，显著提升了图文篡改内容的检测性能。全局与局部对比学习机制的联合引入，有效实现了全局语义对齐与局部篡改区域精确定位之间的

表11 Twitter和Weibo数据集上跨模态门控融合模块消融实验结果

Table 11 Ablation study results of cross-modal gated fusion module on the Twitter and Weibo datasets

/%

| 方法 | Twitter数据集 | | | Weibo数据集 | | |
|---|---|---|---|---|---|---|
| | AUC | ACC | EER↓ | AUC | ACC | EER↓ |
| w/o CGFM | 73.41 | 61.96 | 30.04 | 58.18 | 52.47 | 43.11 |
| 本文 | **76.25** | **65.32** | **28.22** | **60.86** | **56.65** | **41.02** |

注：加粗字体表示各列最优结果。w/o表示消去该组件。↓表示值越小越好。

表12 性能评估结果

Table 12 Performance evaluation results

| 方法 | Params/M↓ | FLOPs/G↓ | AUC/% |
|---|---|---|---|
| VLP-GF | 223.36 | 35.68 | 92.84 |
| HAMMER | **212.18** | **32.43** | 92.70 |
| 本文 | 242.43 | 35.85 | **93.23** |

注：加粗字体表示各列最优结果，↓表示值越小越好。

平衡。模型创新性地设计了伪造感知交互模块与跨模态门控融合模块，前者通过多尺度特征提取与频率域特征融合增强了不同粒度篡改的定位能力，后者通过动态权重分配策略提升了跨模态信息融合的精度与鲁棒性。实验结果表明，模型在4项核心子任务中达到现有最优方法的性能，充分验证了模型在多模态深度伪造检测方面的优势。未来研究将着重优化跨模态融合策略，并探索更复杂多元的多模态数据融合方案。由于使用了大语言模型，本文方法的模型参数量和计算复杂度较高，在未来工作中将尝试采用量化和蒸馏以优化效率，以进一步提升模型在复杂篡改场景下的性能表现。

## 参考文献(References)

Amoroso R, Morelli D, Cornia M, Baraldi L, Del Bimbo A and Cucchiara R. 2024. Parents and children: distinguishing multimodal deepfakes from natural images. ACM Transactions on Multimedia Computing, Communications and Applications, 21(1): #11 [DOI: 10.1145/3665497]

Boididou C, Andreadou K, Papadopoulos S, Nguyen D T D, Boato G, Riegler M, et al. 2015. Verifying multimedia use at MediaEval 2015//Proceedings of the MediaEval 2015 Workshop. Wurzen, Germany: CEUR-WS: 1-6

Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA: ACL: 4171-4186 [DOI: 10.18653/v1/N19-1423]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. 2021. An image is worth 16 × 16 words: transformers for image recognition at scale [EB/OL]. [2025-09-10]. https://arxiv.org/pdf/2010.11929.pdf

Galdi C, Panariello M, Todisco M and Evans N. 2024. 2D-malafide: Adversarial attacks against face deepfake detection systems//Proceedings of 2024 International Conference of the Biometrics Special Interest Group (BIOSIG). Darmstadt, Germany: IEEE: 1-7 [DOI: 10.1109/BIOSIG61931.2024.10786754]

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. 2014. Generative adversarial nets//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: ACM: 2672-2680

Jin Z W, Cao J, Guo H, Zhang Y D and Luo J B. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs//Proceedings of the 25th ACM International Conference on Multimedia. Mountain View, USA: ACM: 795-816 [DOI: 10.1145/3123266.3123454]

Kim W, Son B and Kim I. 2021. ViLT: vision-and-language transformer without convolution or region supervision//Proceedings of the 38th International Conference on Machine Learning. [s. l.]: PMLR: 5583-5594

Li Q L, Gao M L, Zhang G S, Zhai W Z, Chen J Y and Jeon G. 2024. Towards multimodal disinformation detection by vision-language knowledge interaction. Information Fusion, 102: #102037 [DOI: 10.1016/j.inffus.2023.102037]

Liu H, Tan Z C, Chen Q, Wei Y C, Zhao Y and Wang J D. 2025. Unified frequency-assisted transformer framework for detecting and grounding multi-modal manipulation. International Journal of Computer Vision, 133(3): 1392-1409 [DOI: 10.1007/s11263-025-02010-2]

Loshchilov I and Hutter F. 2019. Decoupled weight decay regularization//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview.net

Luo Y C, Zhang Y, Yan J C and Liu W. 2021. Generalizing face forgery detection with high-frequency features//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 16312-16321 [DOI: 10.1109/CVPR46437.2021.01605]

Monu and Dhanakshirur R R. 2024. Herd mentality in augmentation-not a good idea! A robust multi-stage approach towards deepfake detection [EB/OL]. [2025-09-10]. https://arxiv.org/pdf/2410.05466.pdf

Paszke A, Gross S and Lerer A. 2017. Automatic differentiation in PyTorch//Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017). LongBeach, USA: Curran Associates Inc.: 1-4

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 8748-8763

Radford A, Wu J, Child R, Luan D, Amodei D and Sutskever I. 2019. Language models are unsupervised multitask learners [EB/OL]. [2025-09-11]. https://cdn. openai. com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Salvi D, Liu H G, Mandelli S, Bestagini P, Zhou W B, Zhang W M, et al. 2023. A robust approach to multimodal deepfake detection. Journal of Imaging, 9(6): #122 [DOI: 10.3390/jimaging9060122]

Shao R, Wu T X and Liu Z W. 2023. Detecting and grounding multimodal media manipulation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 6904-6913 [DOI: 10.1109/CVPR52729.2023.00667]

Sun K, Chen S, Yao T P, Liu H, Sun X S, Ding S H, et al. 2024. DiffusionFake: enhancing generalization in deepfake detection via guided stable diffusion//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: ACM: #3218

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: ACM: 6000-6010

Wang J K, Wu Z X, Ouyang W H, Han X T, Chen J J, Jiang Y G, et al. 2022. M2TR: Multi-modal multi-scale transformers for deepfake detection//Proceedings of 2022 International Conference on Multimedia Retrieval. Newark, USA: Association for Computing Machinery: 615-623 [DOI: 10.1145/3512527.3531415]

Wang S Y, Feng C B, Liu C X and Jin Y S. 2025. Multivariate and soft blending sample-driven image-text alignment for deepfake detection. Journal of Image and Graphics, 30(5): 1334-1345 (王诗雨, 冯才博, 刘春晓, 金逸胜. 2025. 多元软混合样本驱动的图文对齐人脸伪造检测. 中国图象图形学报, 30(5): 1334-1345) [DOI: 10.11834/jig.240252]

Yamada I, Asai A, Shindo H, Takeda H and Matsumoto Y. 2020. LUKE: deep contextualized entity representations with entity-aware self-attention//Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing. [s.l.]: Association for Computational Linguistics: 6442-6454 [DOI: 10.18653/v1/2020.emnlp-main.523]

Zhang G S, Gao M L, Li Q L, Zhai W Z and Jeon G. 2024. Multi-modal generative deepfake detection via visual-language pretraining with gate fusion for cognitive computation. Cognitive Computation, 16(6): 2953-2966 [DOI: 10.1007/s12559-024-10316-x]

Zhang J, Xu P, Liu W J, Guo X X and Sun F. 2025. Negative instance generation for cross-domain facial forgery detection. Journal of Image and Graphics, 30(2): 421-434 (张晶, 许盼, 刘文君, 郭晓萱, 孙芳. 2025. 多样性负实例生成的跨域人脸伪造检测. 中国图象图形学报, 30(2): 421-434) [DOI: 10.11834/jig.240160]

Zhao H Q, Wei T Y, Zhou W B, Zhang W M, Chen D D and Yu N H. 2021. Multi-attentional deepfake detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 2185-2194 [DOI: 10.1109/CVPR46437.2021.00222]

Zhao W C, Lu Y X, Jiao G and Yang Y. 2024. Concentrated reasoning and unified reconstruction for multi-modal media manipulation//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. Seoul, Korea(South): IEEE: 8190-8194 [DOI: 10.1109/ICASSP48485.2024.10447651]

Zou H Q, Shen M, Hu Y C, Chen C, Chng E S and Rajan D. 2024. Cross-modality and within-modality regularization for audio-visual deepfake detection//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. Seoul, Korea (South): IEEE: 4900-4904 [DOI: 10.1109/ICASSP48485.2024.10447248]

## 作者简介

刘凤阳，男，硕士研究生，主要研究方向为多媒体内容安全和图像处理。E-mail：liufy0618@163.com

张玉金，通信作者，男，副教授，主要研究方向为人工智能和模式识别。E-mail：yjzhang@sues.edu.cn

吴飞，男，教授，主要研究方向为智能信息处理、定位技术和机器学习。E-mail：feiwu1@163.com