基于图像篡改感知与多视图融合网络的虚假新闻检测方法

吴飞[1]，邓哲颖[1]，黄英杰[1]，母兴宇[1]，季一木[2]
（1. 南京邮电大学人工智能学院，南京 210023；2. 南京邮电大学计算机学院，南京 210023）

摘要：AI 与图像修改技术的不断革新导致了篡改图像的激增，这对假新闻检测检测带了新的挑战。针对当前主流多模态假新闻检测方法聚焦图文一致性建模，忽视图像本身可能被篡改，导致模型在面对篡改图像时鲁棒性不足的问题，提出了一种基于图像篡改感知与多视图融合网络（Image Tampering Perception and Multi-view Fusion Network, ITPMFN）的假新闻检测方法。该方法包括三个部分：1）多视图特征提取模块：利用 BERT 与 Swin-T 捕获文本与图像的模态特征，并借助 CLIP 提取图文跨模态对齐的语义特征，构建四通道多视图表征；2）基于协同推理的图像篡改感知与可解释性分析生成模块：采用轻量级模型提取低层篡改统计特征，结合统计特征，设计增强提示，指导多模态大模型生成包含篡改对象、操作类型等信息的结构化高层篡改推理解释，从而提供人类可理解的决策依据，并将该语义解释编码为可融合特征用于下游的假新闻判别；3）跨模态交互与融合模块：分别在模态内与模态间层面基于注意力机制充分交互并融合多视图特征，获得更具鉴别性的融合特征，结合篡改推理特征进行多模态假新闻检测。在 Weibo 与 Fakeddit 两个广泛使用公开数据集上的实验表明，所提方法均优于现有主流方法，消融实验进一步验证了各模块的有效性。

关键词：假新闻检测；多视图学习；多模态融合；图像篡改检测；多模态大模型.
中图法分类号：TP391　　**文献标识号：**A　　**文章编号：**

引言

在信息爆炸与社交媒体高度普及的数字时代，假新闻已成为威胁公共舆论安全、社会稳定的重要隐患。近年来，随着多媒体技术的迅猛发展，假新闻的形式已从传统的纯文本内容演变为融合文本、图像等多种信息的多模态新闻。其中，图像篡改（image tampering）作为一种隐蔽而高效的误导手段，正被广泛用于制造具有高度欺骗性的多模态假新闻[1]。例如，通过深度伪造（Deepfake）、局部内容替换或语义不一致的图文拼接，攻击者可轻易构建看似真实但实质虚假的新闻场景，从而误导公众认知、煽动情绪甚至干预政治进程。

当前多模态假新闻检测研究取得了显著进展，研究者普遍意识到仅使用纯文本信息与视觉信息进行假新闻检测具有片面性，同时考虑文本和图像信息已经得到了学术界的关注。Wang 等人[2]提出了 EANN，它引入了一个额外的事件差异鉴别器来获得事件不变特征。Singhal 等人[3]开发了 Spotfake，它利用迁移学习通过从图像和文本中提取和整合多模态特征来检测假新闻。Zhang 等人[4]提出了 BDANN，一种基于 BERT[5]的域自适应网络，用于捕获上下文和语义信息，以增强假新闻检测。Qian 等人[6]提出了 HMCAN，将多层多模态上下文信息结合起来，对联合文本图像表示进行建模，提高了检测精度。Peng 等人[7]提出了 MRML，通过提取模态关系来强化假新闻检测，基于深度度量学习算法，学习到了不同模态的模态内与模态间的深度关系。Qu 等人[8]提出了 QMFND，其将图像和文本特征传递给量子卷积神经网络以获得判别结果，具有良好的表达能力和纠缠能力，而且对量子噪声具有良好的鲁棒性。Chen 等人[9]提出了 CAFE，通过量化文

基金项目：目家自然科学基金（62076139）。
收稿日期：2026-02-09；修订日期：2026-04-28

本与图像之间的跨模态模糊性，自适应地融合单模态特征与跨模态交互特征，显著提升了检测性能。Chen 等人[10]提出了 MFCL，一种融合文本、图像和传播网络三模态信息的假新闻检测方法，通过对比学习缓解对标注数据的依赖，并引入图文匹配（ITM）与自适应传播网络增强（AP）提升跨模态一致性与高阶特征学习。显然，最近的主流方法[9-12]主要聚焦于图文语义一致性建模，然而，近期综述性工作[13-15]普遍意识到图像、视频等模态的篡改与生成是一个关键问题，并认为忽视图像真实性会严重削弱多模态检测的可靠性。因此，上述工作普遍忽视了对图像本身真实性的验证[16]，导致模型在面对精心设计的篡改图像时鲁棒性严重不足。

图像篡改检测旨在检测篡改图像中的伪影和不规则性识别图像中被修改的区域。Dong 等人[17]利用多尺度监督和多视图特征学习来同时捕获图像噪声和边界伪影。Guo 等人[18]采用了多分支特征提取和定位模块的组合，在处理 CNN 合成和编辑的图像中表现突出。Guillaro 等人[19]提出 TruFor 框架，结合多源线索用于图像篡改检测与定位。然而，这些研究都不具备较好的可解释性，此外 AIGC，DeepFake 等新型篡改形式的出现也带来了新的挑战。Xu 等人[20]提出了 Fakeshield 模型，它是首个将多模态大语言模型应用于图像篡改检测的框架，其在考虑到 copy-move 与 AIGC 等多领域篡改形式的基础上，基于通用多模态大模型的理解与认知生成图像篡改检测的推理结果，从而提升检测的准确性与可解释性。但是，Fakeshield 工作没有考虑到通用大模型在图像篡改领域的幻觉问题，如果推理结果本身是幻觉，那么反而会产生错误结果。此外，这些方法[17-20]局限于单模态问题，现有的新闻多包含如图文在内的多个模态，仅依赖单一模态的方法难以应对图文多模态假新闻的欺骗手段。

针对上述问题，本文提出一种基于图像篡改感知与多视图融合网络（Image Tampering Perception and Multi-view Fusion Network，ITPMFN）的假新闻检测方法。首先，为了获得具有鉴别性的多视图多模态特征，本文设计了多视图特征提取模块，基于不同的特征提取器能力的差异获取同一样本中四个视图特征，交互四视图特征以挖掘特征间的互补信息。随后，本文设计了基于协同推理的图像篡改感知与可解释性分析生成模块，基于轻量级模型提取的低层图像篡改统计特征，设计篡改增强提示指导多模态大模型推理得到高层可解释性的篡改推理特征。最后，本文设计了跨模态交互与融合模块，基于跨模态注意力机制促进了来自不同模态的多视图信息的交互，融合多模态与图像篡改表述特征进行假新闻检测。

因此，本文的主要贡献如下：

（1）设计了一个基于协同推理的图像篡改感知与可解释性分析生成模块，基于小模型高效捕获图像中的篡改痕迹，将其作为结构化证据输入多模态大模型，并由后者推理篡改所涉及的具体对象、表现形式及技术手段，提升了假新闻检测可解释性与鲁棒性。

（2）设计了一个跨模态交互与融合模块来执行四个视图之间的跨模态和跨视图融合，充分考虑了不同模态的相关性和重要性。模态内与模态间的多视图特征融合，增强了模态一致性，挖掘了模态间的互补信息。

（3）实验结果表明，所提方法在公开的 Weibo[21]与 Fakeddit[22]多模态假新闻数据集上均取得显著性能提升。

1 基于图像篡改感知与多视图融合网络的多模态假新闻检测方法

1.1 模型框架

本文提出了一个名为基于图像篡改感知与多视图融合网络（ITPMFN）的多模态假新闻检测方法，旨在进行多视图的跨模态交互，捕捉潜在的图像篡改特征。该方法由三个模块组成，如图 1 所示：1）多视图特征提取模块（MVE），2）基于协同推理的图像篡改检测与可解释性分析生成模块（CRG），3）跨模态交互与融合模块（CIF）。具体来说，MVE 模块从模态与语义两个角度出发，提取四视图特征。随后，CRG 模块使用篡改检测模型 TruFor 提取原始图像的篡改特征，将图像与篡改特征共同输入 Qwen2.5-VL 多模态大模型，生成结构化自然语言篡改判断并编码为篡改推理特征。最后，CIF 模块分别在模态内与模态间交互融合四视图特征，结合篡改推理特征，实现对假新闻的鲁棒判别。本文的方法不仅考虑了多模态语义一致性建模，还考虑到了图像真实性建模，增强了模型在复杂伪造场景下的虚假新闻判别能力与可解释性。

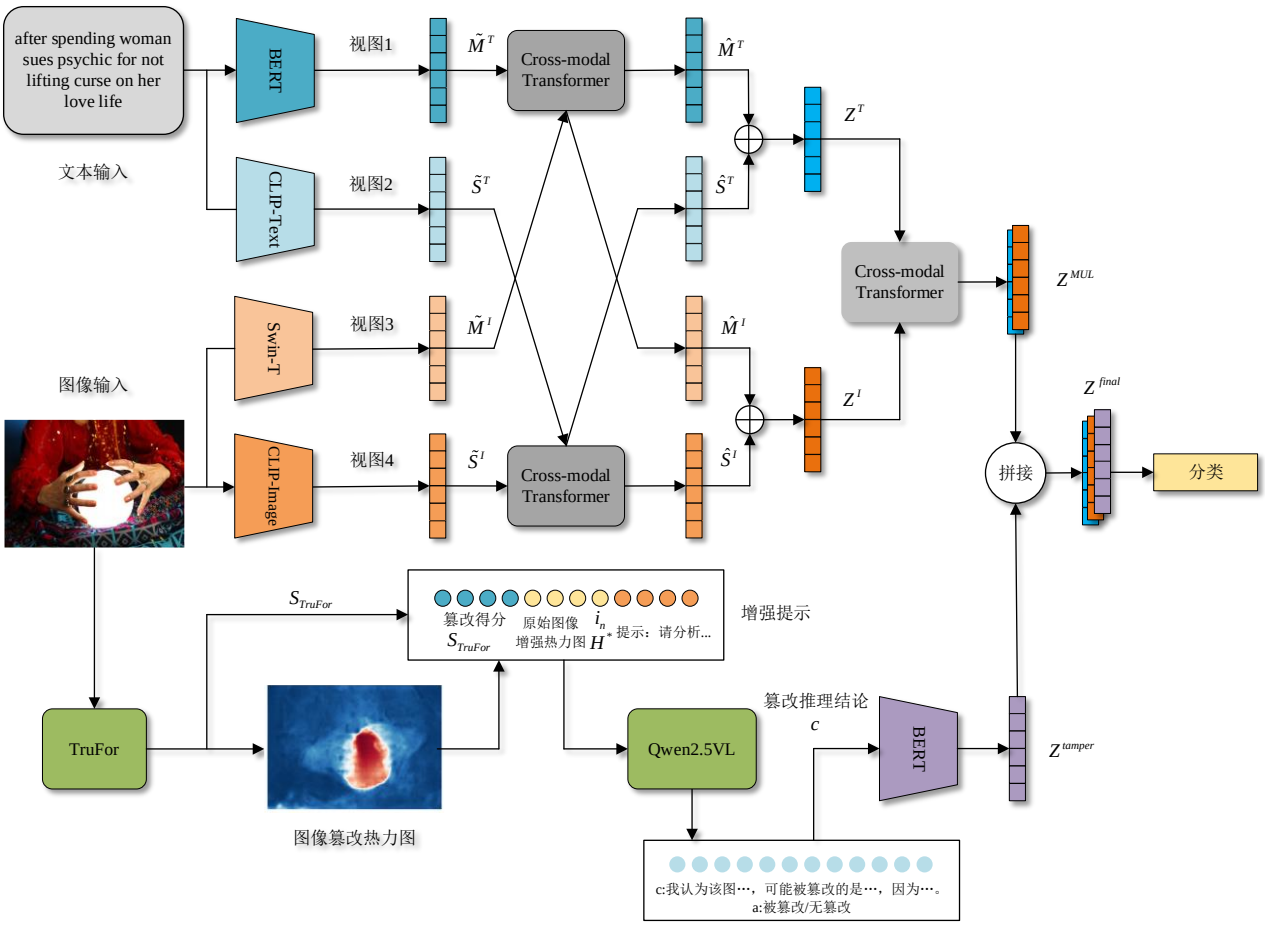

图 1 本文方法总体框架

Fig. 1 The overall framework of the methodology of this paper

1.2 多视图特征提取模块

给定包含文本和图像模态的新闻样本集合$O=\{o_n\}_{n=1}^N$，其中图像模态表示为$I=\{i_n\}_{n=1}^N$，文本模态表示为$T=\{t_n\}_{n=1}^N$，新闻的标签表示为$Y=\{y_n\}_{n=1}^N$。

对于样本o_n，为了充分挖掘其双模态的信息，本文从模态与语义两个角度出发，提取四视图特征：文本模态特征、图像模态特征、文本语义特征与图像语义特征。总体来说，本文使用 BERT、Swin-T[23]、CLIP-Text 和 CLIP-Image 作为四个通道进行四个视图特征的提取。其中，本文使用 BERT 和 Swin-T 来提取样本o_n的模态角度多视图特征，这些特征揭示了样本对应模态的基本细粒度特征。对于语义角度多视图特征，本文使用 CLIP[24]来提取对应特征，该特征揭示了样本o_n模态间的跨模态语义对齐关系。四种特征提取器提供不同视角下对同一样本内容认知的共识与差异，本文基于这种“模态”视图与“语义”视图分离的设计，构建了互补的多视图表征，通过有效整合多视图信息，使模型学习到更鲁棒的特征表示。

具体而言，对于文本模态，本文使用 BERT 提取得到模态特征$M^T=\{m_n^t\}_{n=1}^N\in\square^{d_t^m\times N}$，并使用 CLIP-Text 提取语义特征$S^T=\{s_n^t\}_{n=1}^N\in\square^{d_t^s\times N}$，其中$d_t^m$与$d_t^s$分别表示$m_n^t$和$s_n^t$的维度。对于图像模态，本文通过 Swin-T 提取得到模态特征M^I，其中$M^I=\{m_n^i\}_{n=1}^N\in\square^{d_i^m\times N}$，并使用 CLIP-Image 提取得到语义特征$S^I=\{s_n^i\}_{n=1}^N\in\square^{d_i^s\times N}$，其中$d_i^m$与$d_i^s$分别表示$m_n^i$和$s_n^i$的维度。对于这些不同视图的特征，他们都通过两层多层感知机（MLP）投影到相同的维度。因此，本文可以得到经过编码的文本模态特征$\tilde{M}^T=\{\tilde{m}_n^t\}_{n=1}^N\in\square^{d_k\times N}$、经过编码的文本语义特征$\tilde{S}^T=\{\tilde{s}_n^t\}_{n=1}^N\in\square^{d_k\times N}$、经过编码的图像模态特征$\tilde{M}^I=\{\tilde{m}_n^i\}_{n=1}^N\in\square^{d_k\times N}$与经过编码的图像语义特征$\tilde{S}^I=\{\tilde{s}_n^i\}_{n=1}^N\in\square^{d_k\times N}$，其中$d_k$表示特征的维度。最后，本文定义$Z=\{z_n\}_{n=1}^N\in\{\tilde{M}^T,\tilde{S}^T,\tilde{M}^I,\tilde{S}^I\}$为样本的多视图特征。

在面对复杂的假新闻时，仅依赖模态视图特征可能忽略图文间的深层语义矛盾；而仅依赖语义视图特征又可能丢失单模态内的关键信息。本模块显式地分离并融合这两种视图，基于四种特征提取器从原始新闻中提取出模态与语义层面的互补表征，构建了多视图、多粒度的四视图特征，不仅增强了单模态内部的表征能力，还为后续跨模态交互与融合阶段提供了更全面、更具鲁棒性的特征表示。

1.3 基于协同推理的图像篡改感知与可解释性分析生成模块

为应对当前图像篡改技术日益复杂化所带来的假新闻检测挑战与多模态大模型在图像篡改分析中可能出现的幻觉问题，本文受 X2-DFD[25]的启发，将轻量级模型的检测结果以辅助信息的形式嵌入提示中，并设计了一个双重保险机制，用以引导大语言模型进行更准确的分析与可解释性推理。具体来说，本文提出一种基于协同推理的图像篡改感知与可解释性分析生成模块。该模块从低层统计异常检测扩展到高层语义可解释性推理，该模块统由两个核心阶段构成：（1）基于轻量级模型的低层篡改线索提取，使用领域中

专业的轻量级模型作为第一重保险，提取图像的篡改热力图和全局置信度得分作为硬性篡改先验知识；（2）基于多模态大语言模型驱动的高层篡改解释生成，设计了结构化输入输出模板引导作为第二重保险，要求 MLLM 必须参考所提供的热力图和得分先验知识推理结果，同时回答必须包含具体的篡改对象、原因和技术手段。两个阶段极大地约束了大模型自由发挥的空间，有效降低了产生与图像内容无关的幻觉性解释的风险，生成了准确的，可解释性的篡改推理结论。

1.3.1 基于轻量级模型的低层篡改线索提取

本阶段旨在从输入图像中提取低层篡改特征，生成空间定位热力图与全局置信度得分的篡改特征，为后续语义推理提供初始线索。本文使用 TruFor 模型进行低层篡改特征的提取，TruFor 模型是一种基于 Transformer 的融合架构，它结合了 RGB 图像并通过 Noiseprint++噪声敏感指纹提取器来识别各种局部图像修改，具体流程如下：

给定待检测的原始图像 i_n，将其输送至 TruFor 模型，以获得两项低层统计特征输出。该过程可表示为：

$$H_n,\ S_n^{TruFor} = TruFor(i_n), n = 1,2,\ldots,N \quad (1)$$

其中 H_n 是局部篡改响应热力图，高响应区域指示潜在篡改位置。S_n^{TruFor} 是全局篡改置信度得分，它反映了图像整体被篡改的可能性，满足 $S_n^{TruFor} \in [0,1]$。

1.3.2 基于多模态大语言模型驱动的篡改解释生成

本阶段旨在融合上一阶段轻量级模型的检测结果，构建增强型多模态提示，并以此引导多模态大语言模型（MLLM）进行深层次语义推理，从而生成结构化的且具有可解释性的图像篡改分析结论。特别的，本文使用 Qwen2.5-VL-7B-Instruct[26]作为推理模型，Qwen2.5-VL 模型作为开源模型，具有良好的可复现性，同时其训练过程中包含大规模的多语言图文对，拥有极强的视觉定位、文字理解能力与结构化输出能力。

具体来说，本文首先设计了结构化的提示模板 $P(i_n)$，$P(i_n)$ 包括：（1）原始图像 i_n，（2）结构化提示：“请结合视觉信息[图像]，回答以下问题：图像是否可能被篡改、最可能被篡改的位置或对象、判断存在篡改的原因与具体的篡改手段。回答结论严格参考所给[回答模板]回答。”

表 1 Qwen2.5-VL 篡改增强提示模板

Table 1 Qwen2.5-VL tampering enhancement prompt template

增强提示模板	这是一张新闻的原始图像[图像 1]与对应的篡改检测热力图[图像 2]（红色区域表示高篡改概率）及全局篡改得分[得分]（范围 0~1，越接近 1 表示越可能被篡改）。 请结合 TruFor 模型得到的视觉信息与篡改得分，经过你自己的分析，思考以下问题： 1. 明确是否可能被篡改（是或否，注意不要将图像模糊与篡改混为一谈） 2. 最可能被篡改的目标（具体物体或位置） 3. 篡改的原因（违反物理定律，边缘异常，换脸等） 4. 篡改的手段（多图拼接，复制移动等） 综合以上信息，最终回答一个简洁的结论句，参考如下： 当你认为有篡改时，格式为： ＞“我认为该图有篡改，可能被篡改的物体是[目标名称]，因其[篡改原因]，疑似通过[篡改手段]添加或修改。” 当你认为无篡改时，格式为： ＞“我认为该图无篡改，这张图片是一张[描述图像]的图像。”

随后，本文将全局篡改置信度得分 S_n^{TruFor} 与局部篡改响应热力图 H_n 作为篡改先验知识加入提示模板得到增强提示 $P'(i_n, S_n^{TruFor}, H_n)$，部分增强提示如表 1 所示。基于增强提示 $P'(i_n, S_n^{TruFor}, H_n)$ 的引导，本文将原始图像 i_n 与篡改响应热力图 H_n 作为双视觉输入，送入多模态大语言模型 Qwen2.5-VL-7B-Instruct 中以得到结构化篡改解释响应。通过此机制，Qwen2.5-VL-7B-Instruct 能够协同感知像素级异常与全局置信度，避免对低置信度样本图像产生过度解释。那么最终的输出结构化响应元组可以表示为：

$$R_n = \{a_n, o_n, r_n, m_n, c_n\} = MLLM(P'(i_n, S_n^{TruFor}, H_n)) \tag{2}$$

其中 a_n 为图像篡改判断，o_n 为最可疑篡改区域或对象，r_n 与 m_n 分别为篡改原因与技术手段，c_n 为结构化的文本篡改推理结论。

最后，本文将篡改推理结论 c_n 送入 BERT 提取特征，随后送入两层 MLP 投影到与多视图特征的相同维度，获得篡改推理特征 Z^{tamper}。该阶段基于大小模型协同运作，生成了可靠篡改推理结论，为下游任务提供了高质量篡改推理特征，提高了假新闻检测的可解释性。

1.4 跨模态交互与融合模块

为了有效地整合 2.2 节中得到的多视图特征 Z，本文使用 Cross-Transformer 来实现多视图特征 Z 之间的交互，以便于后续的多模态融合。本文的 Cross-Transformer 为双向跨模态对齐，同时构造图像查询到文本键值对的组合与文本查询到图像键值对的组合，具体如下所示：

$$\begin{aligned} \hat{M}^T, \hat{M}^I &= Cross-Transformer(\tilde{M}^T, \tilde{M}^I, \theta_M) \\ \hat{S}^T, \hat{S}^I &= Cross-Transformer(\tilde{S}^T, \tilde{S}^I, \theta_S) \end{aligned} \tag{3}$$

其中 θ_M 与 θ_S 是不同多视图 Cross-Transformer 的参数，$\hat{M}^T$ 和 $\hat{M}^I$ 是交互后模态视图特征，$\tilde{S}^T$ 和 $\tilde{S}^I$ 是交互后语义视图特征。基于交互后各视图特征，本文使用注意力机制来获得模态内的多视图融合特征：

$$\begin{aligned}\sigma^T &= f(\hat{M}^T;\theta_A)\\ \sigma^I &= f(\hat{M}^I;\theta_A)\\ \varsigma^T &= f(\hat{S}^T;\theta_B)\\ \varsigma^I &= f(\hat{S}^I;\theta_B)\end{aligned} \tag{4}$$

其中 θ_A 与 θ_B 是注意力网络的参数，σ^T 与 σ^I 是交互后模态视图特征 $\hat{M}^T$ 和 $\hat{M}^I$ 的注意力系数，ς^T 与 ς^I 是语义视图特征 $\tilde{S}^T$ 和 $\tilde{S}^I$ 的注意力系数。相应模态视图融合特征可以通过以下方式获得：

$$\begin{aligned}Z^T &= \sigma^T\hat{M}^T + \varsigma^T\hat{S}^T\\ Z^I &= \sigma^I\hat{M}^I + \varsigma^I\hat{S}^I\end{aligned} \tag{5}$$

其中 Z^T 和 Z^I 分别表示在模态内融合得到的文本和图像的输出特征。

随后，为了得到多模态的多视图融合特征，本文采用 Cross-Transformer 来交互多视图特征 Z^T 和 Z^I。

$$\hat{Z}^T,\hat{Z}^I = Cross-Transformer(Z^T,Z^I,\hat{\theta}_{MUL}) \tag{6}$$

其中 θ_{MUL} 是 Cross-Transformer 的参数。然后，基于注意力机制融合特征 $\hat{Z}^T$ 与 $\hat{Z}^I$，并获得融合多模态特征 Z^{MUL}，注意力系数的计算方式与公式 4 计算方式一致：

$$\begin{aligned}\delta^T &= f(\hat{Z}^T;\theta_T)\\ \delta^I &= f(\hat{Z}^I;\theta_I)\end{aligned} \tag{7}$$

其中 θ_T 与 θ_I 是注意力网络的参数，那么多模态特征 Z^{MUL} 可以表示为：

$$Z^{MUL} = \left[\delta^T\hat{Z}^T;\delta^I\hat{Z}^I\right] \tag{8}$$

然后，本文将篡改推理特征 Z^{tamper} 与多模态特征 Z^{MUL} 拼接得到最终融合特征 Z^{final}：

$$Z^{final} = \left[Z^{MUL};Z^{tamper}\right] \tag{9}$$

最后，将融合特征 Z^{final} 输入到一个由单层 MLP 与 ReLU 激活函数组成的分类器中以获得预测标签 $\hat{Y} = \{\hat{y}_n\}_{n=1}^N$，并使用交叉熵损失作为分类损失 L_{cls}，那么分类损失可以表示为：

$$L_{cls} = y_n\log(\hat{y}_n) + (1-y_n)\log(1-\hat{y}_n) \tag{10}$$

2 实验及结果分析

2.1 实验数据集

（1）Weibo 数据集[21]：其包括 3749 条假新闻和 3783 条真实新闻用于训练，1000 条虚假新闻和 996 条真实新闻用于测试。在实验中，本文去除了重复和低质量的图像，以确保数据集的整体质量。

（2）Fakeddit 数据集[22]：Fakeddit 原始训练集有 593187 个图像文本对，原始测试集有 59319 个图像

文本对。在实验中，本论文依据工作[27]的实验设置在 Fakeddit 的训练集中随机抽取了 30000 个图像文本对作为本文的训练集；在测试集中随机抽取了 10000 个图像文本对作为本文的测试集。

2.2 实验细节

本文使用准确率（Accuracy, Acc）、精确率（Precision, P）、召回率（Recall, R）和 F1 分数（F1）来评估 ITPMFN 的性能。

在实验中，批大小设置为 128，并使用初始学习率为 0.0001 的 Adam 优化器进行 100 个 epoch 的模型训练。本文使用预训练的 TurFor 模型与预训练的 Qwen2.5-VL-7B-Instruct 多模态大模型生成对篡改图像的推理与解释，所有的预训练模型都是冻结的，没有进行微调。本文所有代码都是基于 PyTorch 实现的，并在一张 NVIDIA RTX A6000 上运行。

2.3 对比实验与结果

为了验证本文提出方法的有效性，本文将 ITPMFN 方法与 EANN[2]、SpotFake[3]、BDANN[4]、HMCAN[6]、CAFE[9]、MRML[7]、QMFND[8]及 MFCL[10]这八个假新闻检测方法在 Weibo 与 Fakeddit 数据集上进行了比较。

表 2 ITPMFN 方法与其他方法在 Weibo 与 Fakeddit 数据集上的性能比较

Table 2 Performance comparison of ITPMFN method and other methods on Weibo and Fakeddit datasets

数据集	方法	Acc	Fake			Real		
			P	R	F1	P	R	F1
Weibo	EANN	0.782	0.827	0.697	0.756	0.753	0.863	0.804
	SpotFake	0.870	0.855	0.892	0.873	0.769	0.807	0.787
	BDANN	0.842	0.830	0.870	0.850	0.850	0.820	0.830
	HMCAN	0.885	0.920	0.845	0.881	0.856	0.926	0.890
	CAFE	0.898	0.922	0.881	0.901	0.874	0.917	0.895
	MRML	0.897	0.896	0.905	0.901	0.898	0.887	0.892
	QMFND	0.869	0.900	0.810	0.850	0.840	0.920	0.880
	MFCL	0.892	0.911	0.887	0.899	0.871	0.899	0.885
	ITPMFN	**0.939**	0.941	0.943	**0.942**	0.937	0.934	**0.935**
Fakeddit	EANN	0.724	0.727	0.719	0.723	0.722	0.729	0.726
	SpotFake	0.819	0.801	0.848	0.824	0.839	0.790	0.813
	BDANN	0.812	0.836	0.776	0.805	0.791	0.847	0.818
	HMCAN	0.881	0.880	0.882	0.881	0.882	0.880	0.881
	CAFE	0.912	0.946	0.886	0.916	0.878	0.942	0.909
	MRML	0.840	0.819	0.874	0.846	0.865	0.807	0.835
	QMFND	0.942	0.930	0.950	0.940	0.945	0.955	**0.950**
	MFCL	0.938	0.935	0.942	0.938	0.942	0.935	0.939
	ITPMFN	**0.947**	0.966	0.925	**0.945**	0.928	0.968	0.948

在表 2 中，其展示了 ITPMFN 模型和 8 个基线方法在 Weibo 与 Fakeddit 数据集上的检测性能，最佳结果用粗体标出。可以观察到，ITPMFN 方法在两个真实世界的数据集上优于其他方法。

ITPMFN 方法性能改进的原因在于以下两点：1）基于协同推理的图像篡改感知与可解释性分析生成模块高效捕获了图像中的篡改痕迹，基于高质量的先验信息与多模态大模型的推理能力挖掘出了图像中具体的篡改区域及可能的篡改目标、潜在动机与技术手段，额外的篡改信息提升了假新闻检测可解释性与鲁棒性，2）多视图特征提取模块及跨模态交互与融合模块的协同运作，基于 4 种多视图特征，深入挖掘了样本图像文本间的语义关联与互补信息，增强了不同视图间的模态内一致性，获得了更具鉴别性的高质量特征。

2.4 消融实验与结果

在本小节中，本文进行 ITPMFN 方法的消融实验。本文基于三大模块，设计了 5 种消融实验：TPMFN-1 去除了基于协同推理的图像篡改感知与可解释性分析生成模块；ITPMFN-2 去除了多视图特征提取模块，仅采用 BERT 提取文本特征与 CLIP 提取图像特征；ITPMFN-3 仅保留了模态视图的特征；ITPMFN-4 仅保留了语义视图的特征；ITPMFN-4 仅去除了跨模态交互与融合模块，直接拼接多视图特征。表 3 详细展示了每个变体的模型结构，而表 4 汇总了这些 ITPMFN 变体在 Weibo 与 Fakeddit 数据集上的结果。

根据表 4，可以看到五个变体在两个数据集上的表现均劣于完整 ITPMFN 版本。这一现象表明：（1）从语义与模态多视图角度提取与交互多模态特征对虚假新闻检测任务是有益的，（2）对图像进行篡改感知以获取额外的篡改信息也有助于提升假新闻检测的性能。

表 3 消融实验设置

Table 3 The experimental setup for ablation

模型名称	模型结构		
	CRG 模块	MVE 模块	CIF 模块
ITPMFN	保留	保留	保留
ITPMFN-1	消融	保留	保留
ITPMFN-2	保留	消融	保留
ITPMFN-3	保留	保留模态视图	保留
ITPMFN-4	保留	保留语义视图	保留
ITPMFN-5	保留	保留	消融

表 4 ITPMFN 方法在 Weibo 与 Fakeddit 数据集上的消融实验

Table 4 The ablation experiment of ITPMFN method on Weibo and Fakeddit datasets

数据集	方法	Acc	P	R	F1
Weibo	ITPMFN	**0.939**	**0.939**	**0.939**	**0.939**
	ITPMFN-1	0.926	0.927	0.928	0.926
	ITPMFN-2	0.922	0.922	0.922	0.922
	ITPMFN-3	0.872	0.871	0.872	0.872
	ITPMFN-4	0.923	0.925	0.922	0.923
	ITPMFN-5	0.884	0.884	0.884	0.884
Fakeddit	ITPMFN	**0.947**	**0.947**	**0.947**	**0.946**
	ITPMFN-1	0.932	0.933	0.932	0.932
	ITPMFN-2	0.939	0.940	0.939	0.939
	ITPMFN-3	0.924	0.924	0.924	0.924
	ITPMFN-4	0.916	0.916	0.916	0.916
	ITPMFN-5	0.932	0.932	0.932	0.932

2.5 TSNE 可视化

为了更直观地观察分类效果，本文使用 TSNE 在 Weibo 数据集上可视化训练前后的测试集的特征分布。可视化的特征由图像文本与篡改推理三种特征构成，具体来说，假新闻的特征用黄色点表示，而真新闻特征用紫色点表示。在图 2 中，与原始特征相比，训练特征中两类之间的边界变得明显更清晰，这表明 ITPMFN 训练的特征具有更强的判别能力。

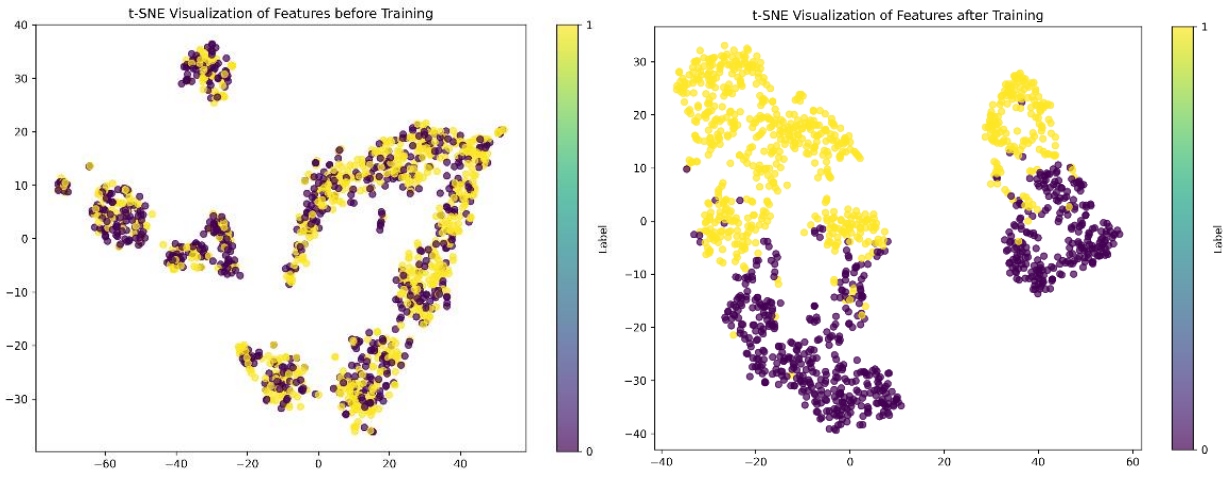

图 2 经过 ITPMFN 训练前后 Weibo 数据集的分布图

Fig.2 Distribution of the Weibo dataset before and after ITPMFN training

2.6 案例分析

为验证所提方法在真实场景下的可解释性与可靠性，本文随机选取了 Fakeddit 数据集中的一个样本进行分析。在图 1 中，可以看到，该新闻的文本为：“after spending woman sues psychic for not lifting curse on her love life”，图像为一名女性手持发光球体。

经过轻量级模型的处理，该图像全局篡改得分为 0.715，对应的篡改响应热力图在人物手中的球体区域呈现显著高响应。随后，基于增强提示引导大模型进行推理，得到如下结论：“我认为该图有篡改，可能被篡改的物体是手中的球，因其边缘异常，疑似通过多图拼接添加。”该结论与热力图定位高度一致，且经人工复核可见：球体边缘存在不自然的锐化与光照不连续现象，极可能为后期合成。此案例表明，引入篡改感知先验可有效引导大模型聚焦真实异常区域，避免依赖语义猜测或产生幻觉，从而生成可信、可验证的篡改推理结论。

2.7 篡改场景验证

为了更直接地检验本文方法在应对图像层面篡改这一关键挑战上的优势，我们进一步在包含丰富图像篡改样本的多模态数据集 DGM4[28]（Detecting and Grounding Multi-Modal Media Manipulation）上进行了补充实验。

具体而言，我们首先从 DGM4 数据集中筛选出标签为 face_swap 和 face_attribute 的样本作为虚假样本，将标签为 orig 的样本作为真实样本。基于此，我们构建了本研究的补充数据集 DGM4-SUB，我们从过滤后的训练集中随机抽取了 5,000 条样本构建训练集，从过滤后的测试集中随机抽取了 1,000 条样本构建测试集。在构建过程中，考虑到类别不平衡的影响，我们严格保证训练集与测试集中的真假样本比例均为 1:1。

我们在 DGM4-SUB 数据集上评估了本文提出的 ITPMFN 方法与当前表现最佳的基线模型 CAFE 的性能。结果如表 4 所示。

表 4　ITPMFN 与 CAFE 在 DGM4-SUB 数据集上的性能比较

Table 4 Performance Comparison of ITPMFN and CAFE on DGM4-SUB Dataset

数据集	方法	Acc	Fake			Real		
			P	R	F1	P	R	F1
DGM4-SUB	CAFE	0.652	0.643	0.684	0.663	0.662	0.620	0.640
	ITPMFN	**0.686**	0.703	0.644	**0.672**	0.672	0.728	**0.699**

从表 4 可以看出，本文方法优于最强基线方法 CAFE，其高出了 0.034 的 Acc，并在 F1 上也具有优势。效果的提高归因于 CAFE 方法专注于量化文本与图像之间的跨模态模糊性，增强单模态特征与跨模态特征的交互与融合，但没有引入图像层面的篡改感知机制，而文本方法的 MVE 模块捕捉到了单模态内部的细粒度信息，CRG 模块有效识别了图像层面的伪造痕迹，提高了存在图像篡改虚假样本的识别能力。因此，本文方法在假新闻存在图像篡改的场景下具有较好的优势。

3 结束语

本文提出了一种基于图像篡改感知与多视图融合网络（ITPMFN）的假新闻检测方法。该方法包括多视图特征提取模块、基于协同推理的图像篡改感知与可解释性分析生成模块、跨模态交互与融合模块。多视图特征提取模块与跨模态交互与融合模块能够有效地提取图文的多视图特征并交互融合，增强模态内一致性并挖掘模态间互补信息。基于协同推理的图像篡改感知与可解释性分析生成模块构建了“低层统计线索—高层语义推理”的协同机制，利用轻量化模型高效定位篡改区域，并引导多模态大语言模型生成结构化、可解释的篡改推理结论，增强了假新闻检测可解释性与准确性。在两个广泛使用的公开数据集上的实验结果、模块的消融实验结果与篡改场景下的实验验证了本文方法的有效性。

参考文献：

[1] 康馨元, 李帆, 赵慧, 等. 基于 CNN-Transformer 混合架构的 AI 生成图像鲁棒检测方法[J]. 数据采集与处理, 2025, 40(5): 1283-1293.
KANG Xinyuan, LI Fan, ZHAO Hui, et al. Robust detection method for AI-generated images based on CNN-transformer hybrid architecture[J]. Journal of Data Acquisition & Processing, 2025, 40(5): 1283-1293.

[2] WANG Y, MA F, JIN Z, et al. EANN: Event adversarial neural networks for multi-modal fake news detection[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London United Kingdom. ACM, 2018: 849-857.

[3] SINGHAL S, SHAH R R, CHAKRABORTY T, et al. SpotFake: A multi-modal framework for fake news detection[C]//2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM). Singapore: IEEE, 2019: 39-47.

[4] ZHANG T, WANG D, CHEN H, et al. BDANN: BERT-based domain adaptation neural network for multi-modal fake news detection[C]//2020 International Joint Conference on Neural Networks (IJCNN). Glasgow, UK. IEEE, 2020: 1-8.

[5] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, MinnesotaAssociation for Computational Linguistics, 2019: 4171-4186.

[6] QIAN S, WANG J, HU J, et al. Hierarchical multi-modal contextual attention network for fake news detection[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. Virtual Event, Canada. ACM, 2021: 153-162.

[7] CHEN Y, LI D, ZHANG P, et al. Cross-modal ambiguity learning for multimodal fake news detection[C]//Proceedings of the ACM Web Conference 2022. Virtual Event, Lyon, France. ACM, 2022: 2897-2905.

[8] PENG L, JIAN S, LI D, et al. MRML: Multimodal rumor detection by deep metric learning[C]//ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece. IEEE, 2023: 1-5.

[9] QU Z, MENG Y, MUHAMMAD G, et al. QMFND: A quantum multimodal fusion-based fake news detection model for social media[J]. Information Fusion, 2024, 104: 102172.

[10] CHEN H, WANG H, LIU Z, et al. Multi-modal robustness fake news detection with cross-modal and propagation network contrastive learning[J]. Knowledge-Based Systems, 2025, 309: 112800.

[11] YU K, JIAO S, MA Z. Fake news detection based on BERT multi-domain and multi-modal fusion network[J].

Computer Vision and Image Understanding, 2025, 252: 104301.
[12] QIAO J, LI X, GAO C, et al. Improving multimodal fake news detection by leveraging cross-modal content correlation[J]. Inf Process Manag, 2025, 62: 104120.
[13] WANG J, LI Z, ZHANG C, et al. Fighting malicious media data: A survey on tampering detection and deepfake detection[J]. Proceedings of the IEEE, 2025, 113(3): 287-311.
[14] GULZAR HUSSAIN F, WASIM M, HAMEED S, et al. Fake news detection landscape: Datasets, data modalities, AI approaches, their challenges, and future perspectives[J]. IEEE Access, 2025, 13: 54757-54778.
[15] MALANOWSKA A, MAZURCZYK W, ARAGHI T K, et al. Digital watermarking: A meta-survey and techniques for fake news detection[J]. IEEE Access, 2024, 12: 36311-36345.
[16] CHANG T P, HSIAO T C, CHEN T L, et al. A framework for detecting fake news by identifying fake text messages and forgery images[J]. Enterprise Information Systems, 2025, 19(3/4): 2436495.
[17] 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.
[18] GUO X, LIU X, REN Z, et al. Hierarchical fine-grained image forgery detection and localization[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada. IEEE, 2023: 3155-3165.
[19] GUILLARO F, COZZOLINO D, SUD A, et al. TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada. IEEE, 2023: 20606-20615.
[20] XU Z, ZHANG X, LI R, et al. FakeShield: Explainable image forgery detection and localization *via* multi-modal large language models[EB/OL]. 2024: arXiv: 2410.02761. https://arxiv.org/abs/2410.02761
[21] JIN Z, CAO J, ZHANG Y, et al. News verification by exploiting conflicting social viewpoints in microblogs[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix, Arizona. ACM, 2016: 2972-2978.
[22] NAKAMURA K, LEVY S, WANG W Y. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection[C]//Proceedings of the Twelfth Language Resources and Evaluation Conference. Marseille, FranceAssociation for Computational Linguistics, 2020: 6149-6157.
[23] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada. IEEE, 2022: 9992-10002.
[24] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[EB/OL]. 2021: arXiv: 2103.00020. https://arxiv.org/abs/2103.00020
[25] CHEN Y, YAN Z, CHENG G, et al. X2-DFD: A framework for eXplainable and eXtendable Deepfake Detection[EB/OL]. 2024: arXiv: 2410.06126. https://arxiv.org/abs/2410.06126
[26] Bai S, Chen K, Liu X, et al. Qwen2.5-vl technical report[J]. arXiv preprint arXiv:2502.13923, 2025. https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct
[27] WU F, CHEN S, DENG Z Y, et al. PFBL: Prototype-based fully balanced learning for multimodal fake news detection[J]. IEEE Transactions on Computational Social Systems, 2025, 12(5): 3442-3454.
[28] SHAO R, WU T, LIU Z. Detecting and grounding multi-modal media manipulation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada. IEEE, 2023: 6904-6913.

作者简介：

吴飞（1989-），**通信作者**，男，教授，研究方向：模式识别与机器学习，E-mail: wufei_8888@126

邓哲颖（2001-），男，硕士研究生，研究方向：模式识别与人工智能。

黄英杰（2001-），男，硕士，研究方向：AI 图像检测与篡改检测。

母兴宇(2004-)，女，研究方向:模式识别与机器学习。

季一木（1978-），男，教授，研究方向：模式识别与人工智能。

Image Tampering Perception and Multi-view Fusion Network Based Fake News Detection

WU Fei[1], DENG Zheying[1], HUANG Yingjie[1], MU Xingyu[1], JI Yimu[2]

(1.School of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing 210023, China;2.School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: The rapid advancement of AI and image editing technologies has led to a surge in tampered images, posing new challenges for Fake News Detection (FND). To address the limitation of prevailing multimodal FND methods, which primarily focus on modeling text-image semantic consistency while neglecting the possibility that images themselves may be manipulated, thereby suffering from insufficient robustness against tampered content, this paper proposes a FND approach based on Image Tampering Perception and Multi-view Fusion Network (ITPMFN). The method consists of three components: (1) a multi-view feature extraction and interaction module that employs BERT and Swin-T to capture modality-specific features from text and images, respectively, and leverages CLIP to extract cross-modal aligned semantic features, constructing a four-channel multi-view representation; (2) a collaborative reasoning-based image tampering perception and interpretable analysis generation module, which first uses a lightweight model to extract low-level statistical tampering cues and then designs enhanced prompts based on these cues to guide a multimodal large language model in generating structured, high-level tampering explanations—including manipulated objects and manipulation types—thereby providing human-interpretable decision rationales, whose semantic embeddings are encoded as fusion-ready features for downstream fake news classification; and (3) a cross-modal interaction and fusion module that applies attention mechanisms to thoroughly interact and fuse multi-view features at both intra- and inter-modal levels, yielding more discriminative representations, which are combined with tampering reasoning features for final multimodal fake news detection. Experiments on two widely used public benchmarks, Weibo and Fakeddit, demonstrate that the proposed method consistently outperforms existing state-of-the-art approaches, and ablation studies further validate the effectiveness of each component.

Highlights:

1.This paper proposes a novel Fake News Detection paradigm based on Image Tampering Perception and Multi-view Fusion Network (ITPMFN), breaking through the limitations of traditional methods that solely rely on text-image semantic consistency.

2.This paper designs a “Low-level Statistical Cues to High-level Semantic Reasoning” collaborative mechanism, which achieves highly reliable interpretable analysis and alleviates Large Model hallucinations.

Keywords: fake news detection; multi view learning; multimodal fusion; image tampering detection; multimodal large model

Foundation item：National Natural Science Foundation of China (62076139)