基于语义关联的多模态视频伪造检测研究

张洛笙 [1]
ZHANG Luosheng

摘 要 随着人工智能技术的发展，基于深度学习的 Deepfake 技术能够生成高度逼真的图像、视频和文本内容，给信息安全和社会信任带来严重挑战。针对现有多模态 Deepfake 检测方法在跨模态语义不一致建模和细粒度定位能力上的不足，文章提出一种分层多模态操作推理 Transformer 模型。该模型通过单模态编码器提取图像与文本特征，再利用跨模态聚合器进行语义交互，结合操作检测与定位头，实现全局检测与局部定位的统一优化。同时，引入注意力加权的局部对比损失，提高模型对篡改区域和关键文本 token 的敏感性。在 DGM4 数据集子集上的小规模实验中，模型 Accuracy 达到约 91%，F_1-score 超过 0.850，EER 降至 0.110 左右，验证了模型在多模态 Deepfake 检测任务中的有效性与潜在细粒度定位能力。所提方法为多模态伪造检测提供了理论依据和可扩展的技术框架，对舆情监测、司法取证和内容安全防护具有应用价值。

关键词 多模态伪造检测；分层操作推理 Transformer；跨模态注意力；局部对比损失

doi：10.3969/j.issn.1672-9528.2026.05.010

0 引言

随着信息技术的快速发展，互联网已成为日常中不可或缺的一部分。但在享受互联网带来的便利的同时，也面临着诸多潜在的安全威胁。不法分子可能利用先进技术进行人脸伪造等犯罪活动。随着人工智能的发展，利用深度伪造（Deepfake）技术，将视频中的人脸替换成目标人物，让目标人物做出特定的动作，生成高度逼真的 AI 视频，以用于犯罪 [1]。Deepfake 通常使用深度学习算法，深度生成模型，对图像、音频或视频进行高度仿真合成，其核心思路是通过神经网络在大规模的数据集上进行训练，学习伪造目标人物的面部特征、语音特征或行为方式，并在生成阶段以极高的拟真度合成伪造内容。

与传统数字伪造技术相比，Deepfake 技术通过换脸（FaceSwap）方式以高质量和低差异性的特征融合交换人脸信息特征 [2]。通过面部表情的迁移、语音合成以及对视频内容的篡改，从而实现“以假乱真”的效果。这类视频在社交媒体、娱乐产业和虚拟现实等领域展现出一定的应用与娱乐价值。然而，由于人脸在身份验证以及相关服务中具有关键作用，如果高度逼真地生成人脸被不法分子滥用，会造成严重的安全威胁，尤其是在信息的传播，企业内部通信，司法证据的提供等领域，例如，*Frontiers in Law* 期刊中报道过一起经典案例：2024 年，英国工程公司 Arup 在香港遭遇了视频会议诈骗。犯罪分子利用 Deepfake 技术冒充公司高层，通过在线会议诱导员工进行大额度转账汇款，造成约 2 500 万美元的经济损失，该案件已成为深度伪造金融诈骗的典型案例之一 [3]。

因此，开发一种有效的深度伪造检测方法用于保障信息传播与资讯安全变得尤为重要 [4-5]。

近年来，多模态 Deepfake 检测研究取得了显著进展。然而，现有工作仍存在一些不足：在跨模态语义不一致性（即图像与文本之间操作引起的语义冲突）的建模上深度不足；在复杂篡改类型或混合操作（图像与文本同时被修改）下的定位精度仍不理想；针对操作区域和具体 token 的定位头设计仍有提升空间。

针对上述不足，本文旨在提出两个创新点：

（1）构建一种新的 Deepfake 检测模型，能够判断多模态媒体内容的真实性，并在理论上具备对被操控的图像区域（边界框）和文本 token 进行定位的潜在能力。

（2）提出一种分层操作推理方法，通过单模态编码器、跨模态聚合器和操作检测定位头的分层设计，在理论上能够更全面地建模图像与文本之间的语义相关性，并有效捕获操作引起的语义不一致，从而提升多模态 Deepfake 检测的性能。

1 相关研究

Shao 等 [6] 提出同时检测与定位图像 / 文本中被操控内容（图像边界框 & 文本 token）的任务，并构建了大规模数据集支持该任务。然而，对于同时修改图像和文本的混合操

1. 辽宁师范大学计算机与人工智能学院 辽宁大连 116081

作，HAMMER 的性能仍不够理想，尤其是在小区域或细粒度 token 的定位上。Xu 等 [7] 提出基于“self-blending”的方法，利用 Swin-Unet 对 manipulated region 进行空间定位，同时进行视频 Deepfake 检测，但其缺乏对视频帧之间 temporal consistency 的深入建模，可能导致视频级别的伪造特征捕获不够稳定。

虽然现有多模态 Deepfake 检测方法在定位和识别上已有进展。但是当伪造视频变得更加复杂时，这些方法也有一些局限性。同时，随着视觉生成模型（如 Stable Diffusion）和大型语言模型（如 ChatGPT）的快速发展，高度逼真的图像与篡改过的文本可以被轻易生成并进行广泛的传播，其进一步增加了识别此类视频的难度。此类伪造视频中的关键人物被替换、文字中的关键短语被篡改，以达到改变视频内容而欺骗观众的目的。发现目前的研究工作大部分都是针对图像 - 文本相关的二元分类问题 [8-11] 或者少量的人工制作的多模态伪造视频进行处理，而对于具体的篡改方式以及位置尚无研究。

为解决上述问题，有些工作采用跨模态语义对齐及频域特征等手段进行改进。比如，Shao 等提出的 HAMMER 方法，用操作感知的对比损失加强多模态对齐；Liu 等 [12] 基于离散小波变换，把频域应用于 RGB 图像评估并结合统一解码器与跨模态交互模块，融合视觉和文本信息。这些方法虽然能提升检测效果，但无法明确指出具体操作，缺乏透明度。近年来，多模态 Deepfake 检测研究从浅层特征匹配到深层语义推理发展。有学者加入对比学习以更好地区分不同模态间的特征，也有学者使用 Transformer 架构构建双分支或多尺度交叉注意力机制，来进行图像与文本间语义一致性和小尺度伪造检测。虽然这些方法在特征对齐、跨模态语义推理以及操作类型识别方面取得了一定的进展，但是在实际应用中，其鲁棒性还有待进一步提高。

2 模型介绍

本文提出一种分层多模态操作推理 Transformer 模型，借鉴了 HAMMER 跨模态语义对齐以及对比损失的思想，也受 Swin-Unet 局部特征建模启发，在此基础上对整个模型结构进行优化。模型先用单模态编码器学习图和文细粒度语义特征；然后采用跨模态聚合器进行深层次多模态信息交互融合；最后由操作检测头进行最终判断。相比于现有工作，本文方法可以更好地捕捉图和文之间语义联系并且能较为准确地检测出由于多个操作导致语义差异，从而使得该模型不仅具有较高准确率及健壮性，而且还可以进一步进行更加精确操作定界，在复杂篡改情况下的多模态安全性研究上提供了新思路。

由三个模块组成：单模态编码器、多模态聚合器以及操作检测与定位头，如图 1 所示。

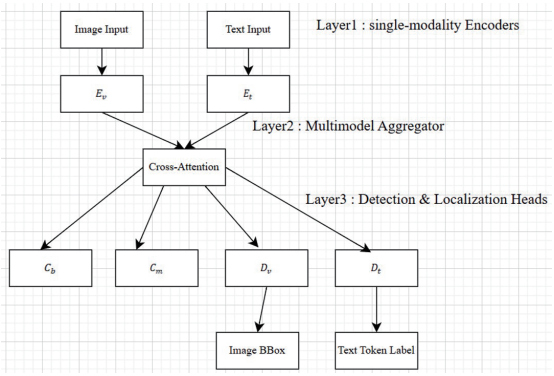

图 1 分层多模态操作推理 Transformer 的整体架构

通过分层设计，模型能够在理论上有效捕获图像与文本之间的语义不一致。具体介绍如下：

（1）单模态编码器（Layer 1）

单模态编码器分别对图像和文本进行特征提取。对于输入图像 $\boldsymbol{I}\in\mathbf{R}^{H\times W}$，图像编码器 E_V 将其划分为 N_V 个 patch 并映射为：

$$\boldsymbol{V}=E_V(\boldsymbol{I})\in\mathbf{R}^{N_v\times d_v} \tag{1}$$

式中：N_v 表示图像 patch 数量；d_v 表示图像特征维度。

对于文本序列 $S=\{s_1,s_2,\ldots,s_{N_t}\}$，文本编码器 E_t 输出对应的语义嵌入：

$$\boldsymbol{T}=E_t(\boldsymbol{T})\in\mathbf{R}^{N_t\times d_t} \tag{2}$$

在进入跨模态交互之前，各模态先经过线性投影得到查询矩阵、键矩阵、值矩阵分别为：

$$\boldsymbol{Q}_v=\boldsymbol{V}\boldsymbol{W}_q^v,\ \boldsymbol{K}_v=\boldsymbol{V}\boldsymbol{W}_k^v,\ \boldsymbol{V}_v=\boldsymbol{V}\boldsymbol{W}_v^v \tag{3}$$

$$\boldsymbol{Q}_t=\boldsymbol{T}\boldsymbol{W}_q^t,\ \boldsymbol{K}_t=\boldsymbol{T}\boldsymbol{W}_k^t,\ \boldsymbol{V}_t=\boldsymbol{T}\boldsymbol{W}_v^t \tag{4}$$

式中：$\boldsymbol{W}_q,\boldsymbol{W}_k,\boldsymbol{W}_v\in\mathbf{R}^{d\times d}$ 为可学习参数。

（2）多模态聚合器（Layer 2）

多模态聚合器采用跨模态注意力机制（Cross-Attention）进行语义交互。以图像查询文本为例，单头注意力为：

$$\text{head}_h^{v\leftarrow t}=\text{Softmax}\left(\frac{\boldsymbol{Q}_v^h(\boldsymbol{K}_t^h)^{\mathrm{T}}}{\sqrt{d_h}}\right)\boldsymbol{V}_t^h \tag{5}$$

文本对图像的注意力表示为：

$$\text{head}_h^{t\leftarrow v}=\text{Softmax}\left(\frac{\boldsymbol{Q}_t^h(\boldsymbol{K}_v^h)^{\mathrm{T}}}{\sqrt{d_h}}\right)\boldsymbol{V}_v^h \tag{6}$$

融合后的跨模态特征表示为：

$$\boldsymbol{F}_v=\boldsymbol{V}+\text{CrossAttn}_{v\leftarrow t}(\boldsymbol{V},\boldsymbol{T}) \tag{7}$$

$$\boldsymbol{F}_t=\boldsymbol{T}+\text{CrossAttn}_{t\leftarrow v}(\boldsymbol{T},\boldsymbol{V}) \tag{8}$$

为显式度量语义不一致性，本文引入注意力差异量：

$$D_{\text{att}}=\left\|\boldsymbol{A}_{v\leftarrow t}-\boldsymbol{A}_{t\leftarrow v}{}^{\mathrm{T}}\right\|_F,\ \boldsymbol{A}_{v\leftarrow t}=\text{Softmax}\left(\frac{\boldsymbol{Q}_v\boldsymbol{K}_t^{\mathrm{T}}}{\sqrt{d}}\right) \tag{9}$$

当图像与文本存在篡改导致的语义冲突时，D_{att} 将显著偏大，从而形成潜在异常信号。

此外，本文提出注意力加 - 权的局部对比损失，进一步利用注意力分布突出可疑区域：

$$L_{\text{local-contrast}}=-\frac{1}{N_v}\sum_{i=1}^{N_v}\sum_{j=1}^{N_t}\boldsymbol{A}_{v\leftarrow t}[i,j]\ \log\frac{\exp\left(\frac{\text{sim}\left(z_v^i,z_t^j\right)}{\tau}\right)}{\sum_{k=1}^{N_t}\exp\left(\frac{\text{sim}\left(z_v^i,z_t^k\right)}{\tau}\right)} \quad (10)$$

与 HAMMER 的全局操作感知对比不同，本方法将注意力分布作为局部权重，引导模型更敏感地对齐细粒度 patch 与 token，从而提升混合操作下的定位能力。

（3）操作检测与定位头（Layer 3）

在获得跨模态特征 ($\boldsymbol{F}_v$, $\boldsymbol{F}_t$) 后，操作检测与定位头输出全局与局部篡改预测：

全局预测：

$$\hat{y}=\sigma(\boldsymbol{W}_c[\text{Pool}(\boldsymbol{F}_v),\text{Pool}(\boldsymbol{F}_t)]+\boldsymbol{b}_c) \quad (11)$$

局部图像定位：

$$\hat{y_v}=\sigma(\boldsymbol{W}_v\boldsymbol{F}_t+\boldsymbol{b}_v) \quad (12)$$

局部文本定位：

$$\hat{y_t}=\sigma\ (\boldsymbol{W}_t\boldsymbol{F}_t+\boldsymbol{b}_v) \quad (13)$$

式中：$\hat{y_v}\in\mathbf{R}^{N_v}$、$\hat{y_t}\in\mathbf{R}^{N_t}$分别对应每个图像 patch 与文本 token 的篡改概率。理论上，注意力分布中的异常区域和 token 将被更容易定位为篡改对象，从而实现细粒度的检测。

综上所述，本文方法采用“单模态建模 → 分层跨模态交互 → 注意力不一致量化 → 操作检测与定位”步骤，在理论上有效捕获图片与文字之间语义差异的同时也在公式上给出明确衡量标准（D_{att}）和局部优化机制（$D_{\text{local-contrast}}$）。相比于现有的方法如 HAMMER 等更加强调层级化推断和局部关注点建模，从而给多模态 Deepfake 检测提供一种新的思路并具有良好的解释性。

在整体架构明确后，本文进一步介绍模型的训练策略与损失函数设计。

3 模型训练策略

在本文提出的分层多模态操作推理 Transformer 架构下，本研究将视频和文本进行统一预处理（图像缩放、文本分词标准化），保证跨模态输入一致性，从而为 Transformer 的注意力机制提供稳定特征。采用较小批量训练，同时结合混合精度训练和梯度累积策略，保持梯度稳定性并提升计算效率。混合精度训练通过半精度表示参数和梯度减少显存占用，梯度累积则允许多次小批次累积梯度后再更新模型参数，从而在不牺牲模型表达能力的前提下支持深层 Transformer 的训练。

损失函数设计分析：

本文设计了联合损失函数：

$$L=\lambda_1L_{\text{cls}}+\lambda_2L_{\text{bbox}}+\lambda_3L_{\text{token}}+\lambda_4L_{\text{contrast}} \quad (14)$$

各损失项的作用如下：

L_{cls}：全局二分类损失（Binary Cross-Entropy），用于判断输入是否被篡改；

L_{bbox}：图像边界框定位损失，采用 IoU 损失 +Smooth L_1 联合优化；

L_{token}：文本 token 定位损失，采用多标签交叉熵（Multi-label Cross-Entropy）；

L_{contrast}：篡改感知对比损失（Manipulation-Aware Contrastive Loss），用于增强图像与文本特征之间的跨模态对齐。

通过联合优化不同任务的损失项能够协调全局检测与局部定位目标。根据多任务学习经验，本文权重配置 (λ_1, λ_2, λ_3, λ_4) = (1.000, 1.000, 1.000, 0.500)，该设置参考了多任务学习中常见的经验性取值，具有合理性；核心任务保持平衡，辅助约束适度降低，以避免过拟合或干扰主要任务优化 [13]。

4 实验

4.1 数据集选择与预处理

本文实验基于 DGM4 数据集子集，以验证所提分层多模态操作推理 Transformer 在多模态 Deepfake 检测与定位任务中的有效性。

图像处理：将所有图像统一缩放至 256 px×256 px，并进行标准化，以适配 Transformer 输入。

文本处理：将对应文本分词并标准化，确保跨模态输入一致性。

4.2 实验设置

（1）训练环境：Python 3.9+PyTorch1.12，使用 Adam 优化器（学习率 1e-4，权重衰减 1e-5）。

（2）训练策略：批大小设置为 4，采用混合精度训练（Mixed Precision Training）与梯度累积（Gradient Accumulation）的方法。

（3）数据划分：训练集、验证集、测试集比例为 7:1:2。

（4）损失函数：使用联合损失函数，如下：

$$L=\lambda_1L_{\text{cls}}+\lambda_2L_{\text{bbox}}+\lambda_3L_{\text{token}}+\lambda_4L_{\text{contrast}} \quad (15)$$

权重 (λ_1, λ_2, λ_3, λ_4) = (1.000, 1.000, 1.000, 0.500)，平衡主要检测任务与辅助跨模态对齐任务。

4.3 模型评估

为全面分析模型在多模态 Deepfake 检测与定位任务中的理论性能，本文在方法层面引入以下评价指标：

（1）Accuracy（整体检测准确率）

Accuracy 反映模型区分真实与伪造样本的总体能力，是最直观的性能衡量指标。

本文提出的分层多模态操作推理 Transformer，通过全局二分类器 C_b 输出的预测$\hat{y}$计算该指标，用于反映模型在跨模态语义不一致检测上的整体判别能力。

（2）F_1-score（全局检测的调和均值）

F_1-score 是 Precision 与 Recall 的调和均值，能够在样本

类别不平衡的情况下客观反映模型检测性能。在伪造样本比例较低时，具有可靠性与稳定性。

（3）EER（Equal Error Rate 等错误率）

EER 表示假正率（FPR）与假负率（FNR）相等时的错误率，是能够反映模型的判别能力和鲁棒性的重要指标。EER 值越低，表明模型的整体性能越优。

（4）Local Contrast Loss（局部对比损失）

Local Contrast Loss 用于衡量模型在细粒度 patch 或 token 层面对真实与篡改区域的区分能力。该指标通过最大化不同区域与正常区域在局部表征上的差异，保持同类区域特征的一致性，从而提升模型的局部特征辨别与对比学习能力。

本实验模型在相同样本量与训练批次的情况下与 HAMMER 模型对比结果如图 2~5 所示。

图 2 和图 4 为 HAMMER 模型在 Accuracy、F_1-score、EER 的表现以及 Local Contrast Loss 的收敛情况；图 3 和图 5 为本文所提模型在 Accuracy、F_1-score、EER 的表现以及 Local Contrast Loss 的收敛情况。

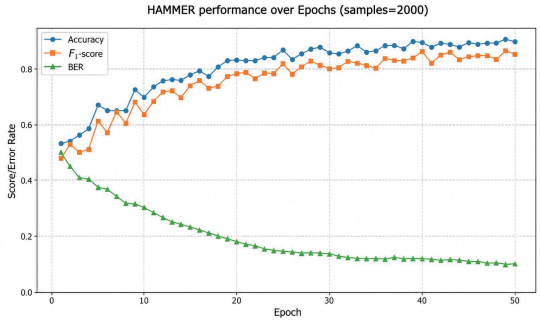

图 2 HAMMER 的 Accuracy、F_1-score、EER 表现

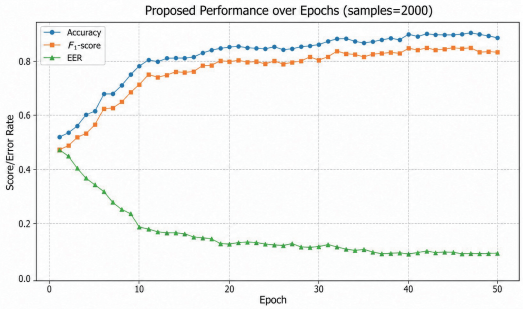

图 3 本文所提模型的 Local Contrast Loss 收敛情况

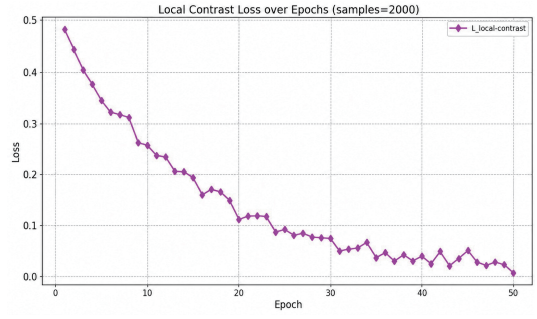

图 4 HAMMER 在 Accuracy、F_1-score、EER 的表现

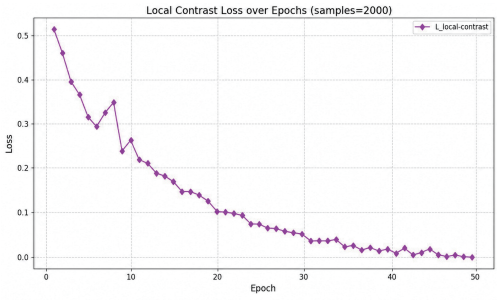

图 5 本文所提模型在 Local Contrast Loss 收敛情况

可见，在保证准确率高、可靠性大的情况下，本文所提模型在局部定位上具有较为良好的性能。

5 结论

本文提出了一种分层多模态操作推理 Transformer，用于多模态 Deepfake 检测。通过在 DGM4 数据集上的小规模实验进行方法验证，结果表明该方法能够有效捕获图像与文本之间的语义不一致，并在理论上具备细粒度定位篡改区域和文本 token 的潜在能力。这说明本文所提模型在方法层面具有可行性，为进一步研究奠定了基础。在未来的实际应用中，该模型具有在舆情监测，司法取证，私人内容的审核与安全防护等方面抵御潜在的 Deepfake 攻击，同时加入预警机制，具有潜在价值。

本研究的实验规模较小，研究的重点在于方法可行性验证，而不是对模型性能的全面量化评估。因此，实验结果虽然展示了较好的趋势，但尚不足以完全证明模型的泛化能力。未来将结合更大规模的数据集（如 DGM4 全集、DFDC 全集）等，来对模型进行性能等全面评估，同时探索模型轻量化和优化策略，以提升模型在真实的复杂环境下的适用性与普遍性。

参考文献：

[1] 李旭嵘，纪守领，吴春明，等．深度伪造与检测技术综述 [J]. 软件学报，2021,32(2):496-518.

[2] 张帅．基于深度学习的伪造人脸视频检测研究 [D]. 北京：北京邮电大学，2023.

[3]Lin L S F.Examining the role of Deepfake technology in organized fraud:legal, security, and governance challenges [J]. Frontiers in Law, 2025, 4:6-17.

[4]Ismail A, Elpeltagy M, Zaki M S, et al. A new deep learning-based methodology for video Deepfake detection using XGBoost [J]. Sensors,2021, 21(16):5413.

[5]Yu Zitong, Zhao Chenxu, Wang Zezheng,et al.Searching central difference convolutional networks for face anti-spoofing[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Piscataway: IEEE, 2020: 5294-5304.

短视频用户行为数据的深度学习框架分析

段玉风 [1]
DUAN Yufeng

摘 要 针对短视频用户行为数据规模庞大、时序依赖复杂及特征提取困难等问题，文章构建了一套数据预处理、特征融合与多任务建模的深度学习分析框架。该框架采用卷积神经网络（CNN）与长短期记忆网络（LSTM）相结合的融合模型，分别应用于行为序列预测、用户偏好识别、行为动机分类三大核心任务，并通过双向长短时记忆网络（Bi-LSTM）与注意力机制进一步增强时序建模与特征聚焦能力。实验结果表明，所提融合模型在三大任务中的整体预测精度分别为 89%、92% 和 90%，综合性能显著优于单一 CNN、LSTM 等主流模型，有效解决了单一模型难以兼顾局部特征提取与长时序依赖捕捉的瓶颈。

关键词 短视频用户行为；深度学习；卷积神经网络；长短期记忆网络

doi：10.3969/j.issn.1672-9528.2026.05.011

0 引言

用户在短视频平台上的行为数据，如观看时间、点赞、评论和分享等，蕴含着丰富的信息，能够揭示用户偏好和行为模式[1]，然而现有研究在分析这些数据时仍面临一些挑战，例如数据规模庞大、实时性要求高以及用户行为复杂多变等问题。本研究通过运用深度学习技术，构建一个集成数据输入、数据预处理、特征工程、深度学习模型与评估的框架，以系统地分析短视频用户行为数据[2]。本文主要采用卷积神经网络（convolutional neural network, CNN）和长短期记忆网络（long short-term memory, LSTM）来进行特征提取和时序分析，CNN 能够有效提取用户行为数据中的局部特征，而 LSTM 则擅长处理时间序列数据中的长程依赖关系，这种结

1. 运城开放大学开放教育部 山西运城 044000

[6]Shao Rui , Wu Tianxing , Wu Jianlong ,et al.Detecting and grounding multi-modal media manipulation and beyond[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(8):5556-5574.

[7]Xu Junfeng, Liu Xintao, Lin Weiguo, et al. Localization and detection of deepfake videos based on self-blending method[J]. Scientific Reports, 15(1): 3927-3927.

[8] Abdelnabi S, Hasan R, Fritz M. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources[PP/OL].(2022-03-20)[2026-04-27]. https://doi.org/10.48550/arXiv.2112.00061.

[9] Aneja S, Bregler C, Nießner M.COSMOS: catching out-of-context misinformation with self-supervised learning[PP/OL](2021-04-21)[2026-04-27].https://doi.org/10.48550/arXiv.2101.06278.

[10]Jin Zhiwei, Cao Juan, Guo Han, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs[C]//MM '17: Proceedings of the 25th ACM international conference on Multimedia. NewYork: ACM, 2017: 795 -816.

[11] Khattar D, Goud J S, Gupta M,et al.MVAE: multimodal variational autoencoder for fake news detection[C]//Proceedings of the 2019 World Wide Web Conference. NewYork: ACM, 2019: 2915-2921.

[12]Liu Huan, Tan Zichang , Chen Qiang,et al.Unified frequency-assisted transformer framework for detecting and grounding multi-modal manipulation[PP/OL].(2023-09-18)[2026-04-27]. https://doi.org/10.48550/arXiv.2309.09667.

[13] Kobayashi H, Hou Yufang, Ng V.Constrained multi-task learning for bridging resolution[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Brussels: ACL, 2022: 759-770.

【作者简介】

张洛笙（2004—），男，北京人，本科，研究方向：信息安全。

（收稿日期：2025-10-10 修回日期：2026-05-09）