Towards General Visual-Linguistic Face Forgery Detection

中文题目：面向通用视觉-语言人脸伪造检测

发布于：CVPR 2025

级别：CCF-A

论文链接：Towards General Visual-Linguistic Face Forgery Detection

论文类别：人脸伪造检测

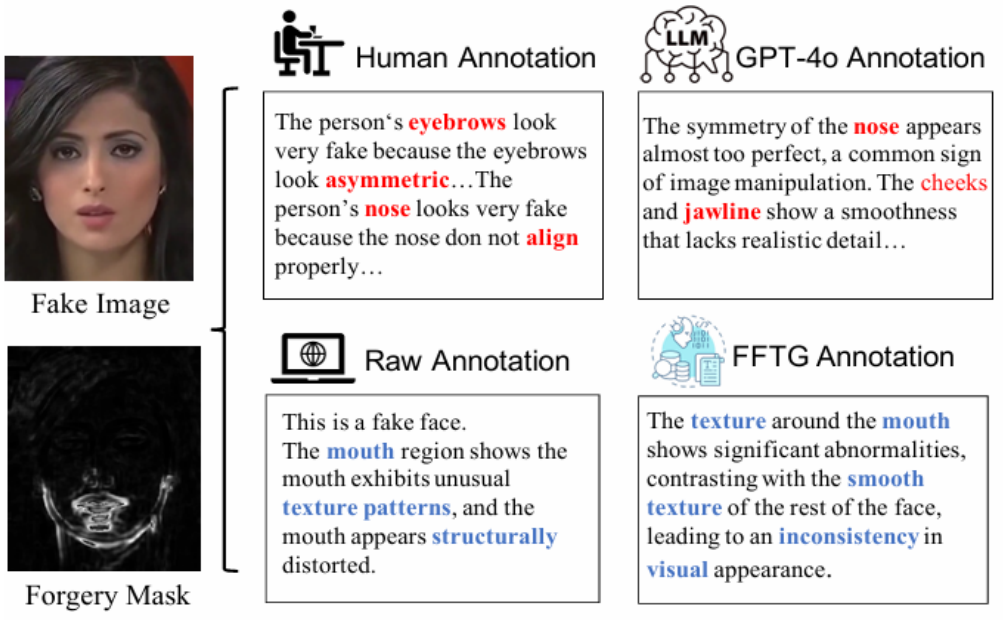
标签：Deepfake，人脸伪造文本生成器，多模态大语言模型，CLIP 微调

本文摘要：人脸篡改技术已取得显著进展，给安全领域和社会信任带来了严峻挑战。近年来的研究表明，借助**多模态模型**能够提升人脸伪造检测的**泛化能力**与**可解释性**。然而，现有的标注方法（无论是人工标注还是直接通过多模态大型语言模型（MLLM）生成标注）往往存在**幻觉**问题，导致文本描述不准确，在高质量伪造内容上表现尤为明显。为解决这一问题，我们提出了人脸伪造文本生成器（Face Forgery Text Generator，FFTG）—— 一种新型标注流水线。该流水线首先借助**伪造掩码**实现初始的伪造区域定位与类型识别，随后通过一套全面的提示策略引导 MLLM 减少幻觉，最终生成准确的文本描述。我们通过两种方式验证了该方法的有效性：**一是采用结合单模态与多模态目标的三分支训练框架对 CLIP 模型进行微调；二是利用我们生成的结构化标注对 MLLM 进行微调。**实验结果表明，我们的方法不仅能生成更准确的标注（伪造区域识别准确率更高），还能在各类人脸伪造检测基准测试中提升模型性能。相关代码已开源，获取地址为：https://github.com/skJack/VLFFD.git。

本文聚焦的问题：

1.**标注幻觉问题**：现有人工或多模态大语言模型（MLLM）生成的伪造检测文本描述，常出现“幻觉”——即描述与真实伪造区域不符，尤其在高质量伪造中会错误地标注未被篡改的区域，导致模型解释性和泛化能力下降。

2.**高质量标注生成问题**：如何利用伪造掩码等视觉线索，构建一个能减少幻觉、生成准确且多样化文本标注的自动化体系，从而提升视觉-语言模型（如 CLIP、LLaVA）在跨域伪造检测中的性能与可解释性。



为什么要引入语言模态？

在伪造检测任务中加入语言，有两个直接的好处：

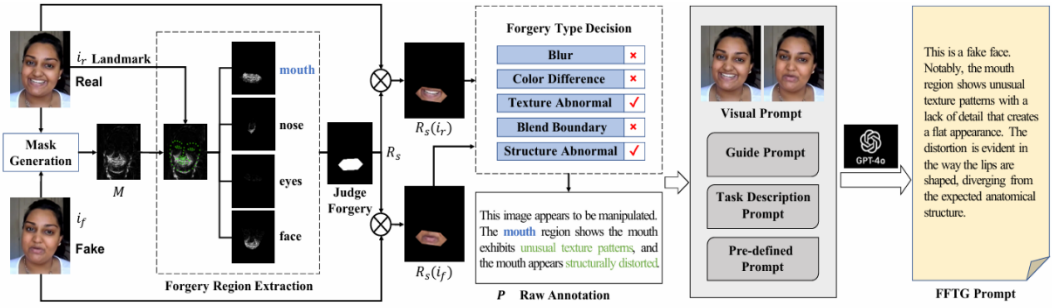
第一，提升可解释性。比起真和假的这种二元黑盒输出，如果模型能进一步说明“假在哪里”“怎么假”，无论是用于分析溯源，还是辅助下游任务，都更有价值；

第二，激活预训练知识。现有的一些视觉 backbone（如 CLIP、LLaVA）等被证明能力已经高于很多纯视觉预训练模型，而这些模型在下游任务的潜在的知识需要语言模态来激活。所以我们希望它们的语言模态不仅能辅助理解图像细节，还能提高模型的迁移能力和泛化表现。

本文提出的方法：

FFTG 伪造文本生成流程

针对上述挑战，研究团队提出了 FFTG（人脸伪造文本生成器），这是一种新颖的标注流程，通过结合伪造掩码指导和结构化提示策略，生成高精度的文本标注。



FFTG 标注流程主要分为两个核心阶段：原始标注生成 (Raw Annotation Generation) 和 标注优化 (Annotation Refinement)。

在这一阶段，FFTG 利用真实图像和对应的伪造图像，通过精确的计算分析生成高准确度的初始标注：

1、掩码生成 (Mask Generation)：

- 通过计算真实图像和伪造图像之间的像素级差异，生成伪造掩码 M
- 掩码值被归一化到 [0,1] 范围，突显操作强度较大的区域

2、伪造区域提取 (Forgery Region Extraction)：

- 基于面部特征点将人脸划分为四个关键区域：嘴部、鼻子、眼睛和整个脸部
- 计算每个区域内掩码 M 的平均值，并设置阈值 θ 判断该区域是否被篡改
- 形成伪造区域列表，并从中随机选择一个区域进行下一步分析

3、伪造类型判定 (Forgery Type Decision)： 设计了五种典型的伪造类型判断标准：

- 颜色差异 (Color Difference)：通过 Lab 色彩空间中的均值和方差差异检测
- 模糊 (Blur)：使用拉普拉斯算子量化局部模糊程度
- 结构异常 (Structure Abnormal)：使用 SSIM 指数衡量结构变形
- 纹理异常 (Texture Abnormal)：通过灰度共生矩阵 (GLCM) 对比度衡量纹理清晰度
- 边界融合 (Blend Boundary)：分析融合边界的梯度变化、边缘过渡和频域特征

4、自然语言描述转换：

- 将识别出的伪造区域和类型转换为自然语言表达
- 如”Texture Abnormal”转换为”lacks natural texture”，”Color Difference”转换为”has inconsistent colors”

此阶段生成的原始标注虽然结构相对固定，但准确度极高，为后续优化提供了可靠基础。

第二阶段：标注优化

为增加标注的多样性和自然流畅性，FFTG 使用多模态大语言模型（如 GPT-4o-mini）进行标注优化，同时设计了全面的提示策略防止幻觉：

1、视觉提示 (Visual Prompt)：

- 将真实和伪造人脸图像作为配对输入提供给大模型
- 这种对比方式使模型能通过直接比较识别伪造痕迹，减少幻觉
- 保持伪造检测视角，避免生成与伪造无关的描述

2、指导提示 (Guide Prompt)：

- 将前一阶段生成的原始标注作为指导提供给大模型
- 附带详细解释每种伪造类型的判定标准（如纹理异常是如何通过 GLCM 分析确定的）
- 强化技术依据，减少主观臆断

3、任务描述提示 (Task Description Prompt)：

- 设定专家级伪造检测任务情境
- 提供分析视觉证据和生成综合描述的具体要求
- 引导模型进行逐步推理

4、预定义提示 (Pre-defined Prompt)：

- 规定输出格式（如 JSON 结构）
- 要求包含特定短语（如”This is a real/fake face”）
- 确保不同样本的标注格式一致

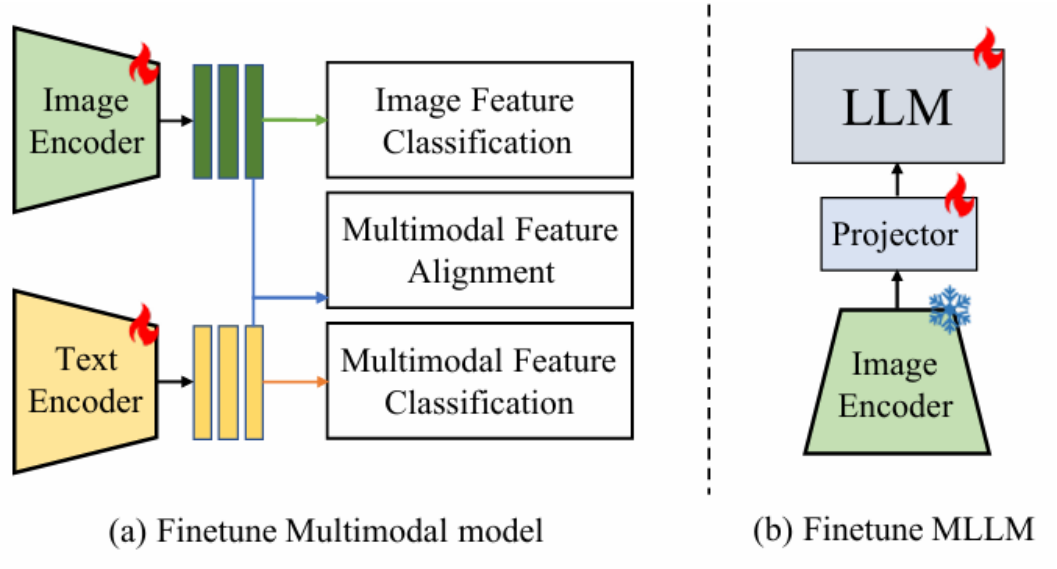
下游微调：双路模型训练策略

有了高质量的图文标注数据，接下来的问题是：如何充分利用这些数据来训练模型？研究团队提出了两种不同的训练策略，分别针对 CLIP 架构和多模态大语言

模型（MLLM），注意本文的目的主要是验证数据的有效性，所以才去了相对简单的微调方式：

CLIP 三分支训练架构

对于 CLIP 这类经典的双塔结构模型，团队设计了一种三分支联合训练框架，如图 4 所示。



这种训练方法结合了单模态和多模态的学习目标：

1、图像特征分类（Image Feature Classification）：直接使用图像编码器提取的特征进行真伪二分类，保证模型在纯视觉输入下的基本检测能力。

2、多模态特征对齐（Multimodal Feature Alignment）：通过对比学习，使图像特征和对应的文本特征在表示空间中对齐，并且激活 CLIP 预训练时获得的跨模态理解能力。

3、多模态特征融合分类（Multimodal Feature Classification）：通过注意力机制融合视觉和文本特征，引导模型学习跨模态的伪造证据整合能力

这三个分支的损失函数共同优化，使模型既能独立运行，又能充分利用文本信息来增强检测能力。

MLLM 微调方法

对于如 LLaVA 这类多模态大语言模型，采用了一种更为直接的微调方法：

MLLM 通常由三部分组成：视觉编码器、对齐投影器和大语言模型。策略是：固定预训练好的视觉编码器参数，专注于微调对齐投影器和大语言模型部分

设计简洁有效的提示模板：”Do you think this image is of a real face or a fake one? Please provide your reasons.”

这种双部分提示不仅引导模型做出二分判断，还要求提供可解释的理由。

实验

标注质量评估

首先，比较了不同标注方法的质量：

Method	Annotation Evaluation			CLIP Evaluation		MLLM Evaluation			
	Precision	Recall	F1	AVG-AUC	AVG-EER	FFpp-ACC	CDF-ACC	Precision	Recall
w/o Text	-	-	-	84.36	20.64	50.13	65.30	10.41	8.10
DD-VQA (Human)	62.46	51.52	52.06	88.25	18.04	73.54	65.60	62.94	53.62
GPT-4o-mini	61.27	44.00	47.18	87.56	19.21	94.84	73.98	58.26	41.85
FFTG	**89.48**	**57.12**	**64.96**	**89.08**	**17.61**	**95.84**	**75.00**	**88.07**	**55.30**

Table 1. Comparison of different annotation approaches. We report precision, recall and F1-score for annotation quality evaluation, AUC and EER for CLIP-based forgery detection and classification accuracy (ACC) and explanation quality (Precision/Recall) for mLLM evaluation on FFpp and Celeb-DF (CDF) datasets.

结果表明，FFTG 在所有指标上都显著优于现有方法。特别是在精度上，FFTG 比人工标注高出 27 个百分点，比直接使用 GPT-4o-mini 高出 28 个百分点，证明了该研究的掩码引导和结构化提示策略能有效减少”幻觉”问题。

跨数据集泛化能力评估

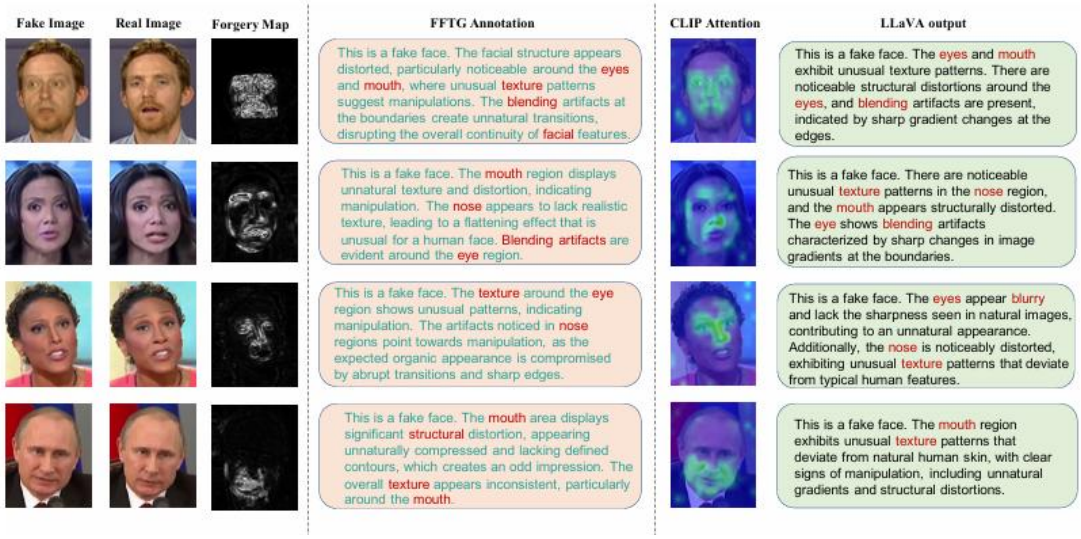
在 FF++数据集上训练模型，并在其他四个未见过的数据集上测试，评估方法的泛化能力：

Method	*FF++*		DFD		DFDC-P		Wild Deepfake		Celeb-DF	
	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER
Xception [9]	99.09	3.77	87.86	21.04	69.80	35.41	66.17	40.14	65.27	38.77
EN-b4 [50]	99.22	3.36	87.37	21.99	70.12	34.54	61.04	45.34	68.52	35.61
Face X-ray [50]	87.40	-	85.60	-	70.00	-	-	-	74.20	-
F3-Net [39]	98.10	3.58	86.10	26.17	72.88	33.38	67.71	40.17	71.21	34.03
MAT [63]	99.27	3.35	87.58	21.73	67.34	38.31	70.15	36.53	70.65	35.83
GFF [37]	98.36	3.85	85.51	25.64	71.58	34.77	66.51	41.52	75.31	32.48
LTW [44]	99.17	3.32	88.56	20.57	74.58	33.81	67.12	39.22	77.14	29.34
LRL [3]	**99.46**	**3.01**	89.24	20.32	76.53	32.41	68.76	37.50	78.26	29.67
DCL [46]	99.30	3.26	91.66	16.63	76.71	31.97	71.14	36.17	82.30	26.53
PCL+I2G [64]	99.11	-	-	-	-	-	-	-	81.80	-
SBI [43]	88.33	20.47	88.13	17.25	76.53	30.22	68.22	38.11	80.76	26.97
UIA-ViT [65]	-	-	94.68	-	75.80	-	-	-	82.41	-
RECCE [1]	99.32	3.38	89.91	19.95	75.88	32.41	67.93	39.82	70.50	35.34
UCF [57]	97.05	-	80.74	-	75.94	-	-	-	75.27	-
CLIP [40]	99.09	3.16	89.03	17.13	78.83	28.95	77.71	30.38	77.16	29.30
Ours	99.16	3.11	**94.81**	**15.22**	**83.21**	**22.43**	**85.10**	**23.65**	**83.15**	**23.66**

Table 2. **Frame-level** cross-database evaluation from FF++(HQ) to DFD, DFDC-P, Wild Deepfake and Celeb-DF in terms of AUC and EER. The FF++ belongs to the intra-domain results while others represent the unseen-domain.

可视化分析

团队对模型的注意力机制进行了可视化分析，进一步验证了 FFTG 的有效性：



总结：

优点

创新性强：首次系统性地提出了一个结合伪造掩码（forgery mask）与多层提示策略的文本生成框架——**Face Forgery Text Generator（FFTG）**，有效缓解了多模态标注中普遍存在的“幻觉”问题。

实验验证充分：论文在五个主流数据集（如 FF++、DFDC-P、Celeb-DF 等）上验证了方法的有效性，实验结果显示 FFTG 显著提升了模型在跨域伪造检测中的泛化能力与解释性。

可解释性提升明显：通过可视化证明，FFTG 生成的文本描述能引导模型关注真实伪造区域，从而提升检测的可解释性。

缺点：

1. **标注生成依赖伪造掩码**：FFTG 需要真实伪造掩码作为输入，这在实际开放场景中难以获得，限制了方法的可应用性。
2. **计算与人工成本较高**：尽管减少了人工标注，但 FFTG 的多阶段处理（掩码生成、特征计算、MLLM 推理）仍需较高计算资源与复杂的工程实现。
3. **多模态模型依赖性强**：FFTG 的性能依赖于多模态大语言模型（如 GPT-4o-mini）的能力，当模型理解力不足或提示策略不匹配时，生成质量可能下降。