双料冠军！TeleAI 登顶 IJCAI 2025 多模态深度鉴伪挑战赛

2025-08-20 09:36 中国电信人工智能研究院（TeleAI）

今年初，法国媒体《巴黎人报》的一则新闻引发全球关注和热议。一法国女子被冒充美国男星“布拉德·皮特”的骗子骗走毕生积蓄。这位自称是“皮特”的人通过 AI 换脸，向该女子发起爱情攻势，并以患病需要治疗为由向其借钱，成功骗走 83 万欧元（约合 629 万元人民币）。

图片来源于网络

随着人工智能技术的迅猛发展，深度伪造（Deepfake）技术通过利用生成对抗网络（GANs）和其他高级算法，能够以惊人的精度合成或篡改音视频内容，使其与真实素材几乎难以区分。这种技术在娱乐和创意领域虽能带来新的应用前景，但同时被越来越多地应用于精准诈骗中，正严重威胁着社会安全与可信度。

域顶级国际会议 IJCAI2025 在加拿大蒙特利尔召开。会议同期举办了一场**主题为“深度伪造检测、定位、可解释性”的研讨会暨挑战赛**，正是为了应对现实中的虚假信息，为可信 AI 生态的构建提供底层技术支撑。

本次挑战赛分为**“图像检测和定位”、“音视频检测和定位”**两个赛道，分别针对图片和音视频中的伪造内容进行精准检测。中国电信人工智能研究院（TeleAI）与来自国内外各地的百余支队伍展开激烈角逐，最终凭借创新的模型训练方法和出色的能力表现脱颖而出，**斩获双赛道第一名。**

在中国电信集团 CTO、首席科学家、中国电信人工智能研究院（TeleAI）院长李学龙教授的指导下，**TeleAI 科研团队针对图像伪造，提出了 LOUPE 模型**，在提升精度的同时，可对未知篡改类型的自适应检测；针对视频伪造，TeleAI 还提出了一种**多模态深度伪造检测方法“ERF-BA-TFD+”**，结合视觉和音频两种信息检测深伪视频中的细微差异。

图像、音视频深度鉴伪技术是 TeleAI 在**“AI 安全与治理”**领域的核心突破之一。**AI 治理**通过对人工智能的研发、应用与管理全流程进行安全约束与正向引导，它既是抵御虚假信息、数据滥用

TeleAI 将 **AI 治理**与智传网（AI Flow）、智能光电（包括具身智能）、智能体结合，形成**“一治+三智”**战略科研布局，目标打通**“AI 驱动的三大空间经济”**，赋能赛博空间、临地空间、广域空间，并打造**安全、可控、可信**的人工智能。

赛道1

图像检测和定位

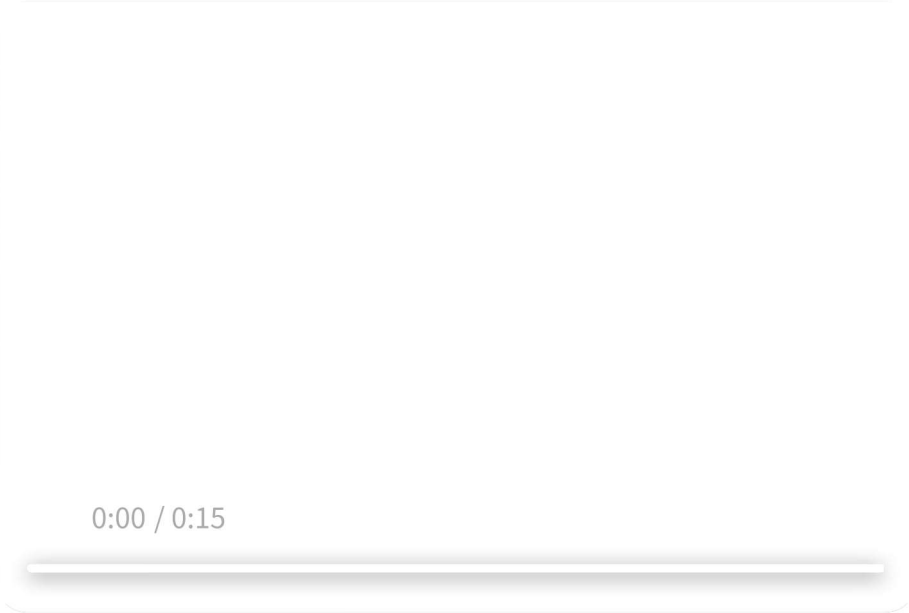
在图像深度鉴伪方面，现有的检测技术在部分场景下能带来较好表现，但仍存在诸多挑战。例如，当面对训练集中未出现的伪造手法时，泛化能力不足，性能显著下降；或由于架构复杂，大量依赖多阶段或多模型组合，导致推理开销大，不利于部署；还有的虽然能判断真伪，但受限于定位精度，难以精准标注伪造区域。

为此，TeleAI 提出了**基于 Mask2Former 的 LOUPE 模型**，通过提出创新的 Patch-aware 分类器、引入条件像素解码器，并设计一套伪标签引导的测试时自适应机制，在保持模型结构简洁的同时，能够实现强泛化、高效率和精准定位。

•**Patch-aware 分类器：**结合全局预测与局部 Patch 预测，不仅能够提升模型的检测准确率，还为后续分割提供了细粒度的伪造

•条件像素解码器：在多尺度特征融合过程中加入条件查询，使得分割结果与图像语义保持高度一致。

•伪标签引导的测试时自适应机制：利用 Patch 级预测作为低分辨率掩膜伪标签，在测试阶段持续优化分割头的表现，从而显著提升跨域鲁棒性。



LOUPE 图像伪造检测系统演示

LOUPE 模型的运行过程分为**“三阶段训练”**和**“一阶段推理”**。

在训练阶段，首先通过图像编码器提取视觉特征。采用 Perception Encoder 作为视觉骨干，将输入图像转化为多尺度视觉特征，为后续的分类和分割任务提供统一的特征输入。

随后，训练分类器。通过冻结编码器，仅训练 Patch-aware 分类器，使用 Poly Focal Loss 缓解类别不平衡问题，并结合 BCE 实现全局-局部预测融合。

再后，训练分割器。基于 Mask2Former 框架，引入条件像素解码器，结合 Tversky Loss 控制精确率与召回率的平衡。

实验结果显示，LOUPE 模型在 DDL 验证集上的表现显著优于多数现有方法。通过消融实验，移除 Patch-aware 分类器或条件像素解码器均会导致性能下降，验证了两者的关键作用。

在本次 IJCAI 2025 “深度伪造检测、定位、可解释性”主题挑战赛中，**TeleAI 凭借 LOUPE 模型获得“图像检测和定位”赛道第一名。**

Table 1: **Leaderboard of the IJCAI 2025 Deepfake Detection and Localization Challenge.** The *overall* score is computed as the average of AUC, F1, and IoU.

Rank	AUC	F1	IoU	Overall
1 (ours)	0.963	0.756	0.819	**0.846**
2	-	-	-	0.8161
3	-	-	-	0.8151
4	-	-	-	0.815
5	-	-	-	0.815

赛道2

音视频检测和定位

随着深度伪造技术的迅速发展和普及，区分真实与篡改的多媒体内容变得愈发困难。这些涉及音视频合成或篡改的攻击，对数字媒体的可信度构成了严重威胁。为应对此挑战，已有研究提出了多种检测方法，但现有系统仍面临三大核心挑战。

第一，检测维度单一。传统的检测方法大多关注单一模态（视频或音频）的伪造痕迹，难以有效应对音视频流同时或分别被篡改的复杂多模态攻击。

第二，场景局限性强。现有的大多数公开数据集主要依赖于短视频片段，这导致模型训练和评估无法反映真实世界中长视频、音

第三，特定伪造失效。部分模型在处理贯穿整个视频的“全伪造”内容时表现不佳，因为缺乏视频内部的真假内容对比，导致检测器难以建立有效的判断基线。

为了应对这些挑战，TeleAI 提出了“**ERF-BA-TFD+”作为新型多模态深度伪造检测模型。**与传统系统不同，该模型通过一种**增强感受野（ERF）与音视频融合的协同框架**，能够有效提升系统在复杂、真实场景下的检测能力。

ERF-BA-TFD+ 模型的核心在于能够同时处理音频和视频特征，并对它们之间的长时序依赖关系进行建模，从而更精准地捕捉真实内容与伪造内容之间的微妙差异。

ERF-BA-TFD+ 音视频伪造鉴别系统演示

TeleAI 通过引入**跨模态重构机制（CRATrans）**和**基于全局证据的推理框架（ERF）**，显著提升了在**处理音视频异步**、**长时程伪造**等挑战性任务中的表现。

•跨重构注意力变换器（CRATrans）：通过强制一个模态重构另一个模态的特征，有效发现音视频流之间的异步性，从而揭示伪造痕迹。

器识别的“全视频伪造”场景。

ERF-BA-TFD+ 模型旨在通过协同分析模态内特征、跨模态关系和全局视频证据，来精准检测和定位伪造内容。

首先，是特征提取，视觉流采用多尺度的 MViTv2 作为骨干网络，以捕捉从细粒度像素到粗粒度运动的各类视觉伪影；音频流则使用自监督模型 BYOL-A，学习丰富的音频语义表示，鲁棒地检测各种听觉异常。

其次，在跨模态异常检测阶段，核心模块 CRATrans 采用一种对抗性的跨重构机制，通过基于 Transformer 的“编码器-解码器”架构，利用一个模态的特征来重构另一个模态。在伪造内容中，由于音视频不一致，重构误差会显著升高，成为可靠的伪造指标。

随后，采用从粗到精的策略进行分层伪造定位。先通过独立的帧级分类器对每个模态进行初步筛选，再通过边界定位模块（Boundary Localization Module）利用提案关系块（PRB）生成精确的边界图，从而确定篡改片段的起止点。

最后，进行全局推理与后处理。引入基于证据的推理框架（ERF）来处理“全视频伪造”的特殊情况，通过分析整个视频检测分数的统计分布来识别那些缺乏明显伪造峰值的全局性伪造。

实验结果显示，ERF-BA-TFD+ 在 LAV-DF 数据集的测试中显著优于多种现有的单模态及多模态检测方法，在多项关键指标上均取得最优成绩。在 AP@0.5 指标上，模型达到了 0.9573，远超 TriDet（0.8633）和 ActionFormer（0.8523）等先进方法。

MDS [25]	0.1278	0.0102	0.0000	0.3788	0.3071	0.3459	0.3215
AGT [26]	0.1785	0.0942	0.0011	0.4315	0.3423	0.2459	0.1671
BSN++ [27]	0.5641	0.3257	0.0021	0.7493	0.7111	0.6498	0.5929
AVFusion [28]	0.6538	0.2389	0.0011	0.6298	0.5926	0.5480	0.5211
BA-TFD [29]	0.7915	0.3857	0.0024	0.6703	0.6418	0.6089	0.5851
TadTR [30]	0.8022	0.6104	**0.0522**	0.7250	0.7250	0.7056	0.6918
ActionFormer [31]	0.8523	0.5905	0.0093	0.7723	0.7723	0.7719	0.7693
TriDet [32]	0.8633	0.7023	0.0305	0.7447	0.7447	0.7446	0.7445
ERF-BA-TFD+ (ours)	**0.9573**	**0.8318**	0.0354	**0.8085**	**0.7971**	**0.7866**	**0.7793**

“ERF-BA-TFD+”模型在 LAV-DF 测试集上的性能对比

在 AR 指标上，模型同样表现出色，证明了其在复杂场景下检测和定位伪造内容的高效性和鲁棒性。通过引入 UMMA 框架和 ERF 模块，模型成功解决了**音频伪造检测**和**长视频伪造检测**中的关键挑战。

在本次 IJCAI 2025 “深度伪造检测、定位、可解释性”主题挑战赛中，**TeleAI 凭借 ERF-BA-TFD+ 模型获得“音视频检测和定位”赛道第一名。**

AP@0.5	AP@0.75	AP@0.95	AR@90	AR@50	AR@20	AR@10
0.9243	0.8050	0.0451	0.8246	0.8121	0.8039	0.7952

“ERF-BA-TFD+”模型在 DDL-AV 测试集上的成绩

上一篇 数据标注新突破！中国电信中标东莞市数据标注产业基础设施项目，助力打造粤港澳大湾区工业数据价值转化中枢

下一篇 35款热门AI能力已开通购买，钜惠来袭