田玉政, 蔡满春. 基于自监督学习的多流融合视频伪造检测方法[J]. 智能计算机与应用,2025,15(12):39-46. DOI:10.20169/j.issn.2095-2163.25060901

基于自监督学习的多流融合视频伪造检测方法

田玉政, 蔡满春
(中国人民公安大学 信息网络安全学院, 北京 100038)

摘 要: 近年来,深度伪造对在线媒体的真实性构成了严峻挑战,引发了公众的担忧。现实世界场景中的伪造视频通常由音频和视觉模态的双重伪造组成,主流的深度伪造检测方法是通过捕获视觉单模态中的伪影来实现的,不同的伪造方法会生成不同的伪影,导致此类方法的性能受限和泛化能力较差;利用多模态信息进行检测的工作并没有充分利用音频信息发掘视觉模态的伪造。为此,本文提出了一种基于自监督学习的多流融合的视频伪造检测方法(MFAVNet),利用音频和视觉模态之间的不一致性进行多模态伪造检测。构建视频、音频、视听 3 个平行分支的特征提取网络,使用自监督预训练的 Audio-Visual HuBERT(AV-HuBERT)模型提取嘴唇区域视觉特征和音频特征;为补充面部区域的整体视觉特征,设计基于 ResNet 架构的视觉-音频特征提取器,并在伪造数据集上进行微调优化,经特征融合,通过多层感知机实现最终分类。在 DeepfakeTIMIT 和 FakeAVCeleb 数据集上的实验结果表明,本文所提出的方法在检测多种伪造技术生成的视频时展现了良好的泛化适应性,在深度伪造检测方面的性能优于许多主流先进的方法。

关键词: 深度伪造;自监督;多模态;特征融合;视觉;音频
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2025)12-0039-08

Multi-stream fusion video deepfake detection method based on self-supervised learning

TIAN Yuzheng, CAI Manchun
(School of Information and Cyber Security, People's Public Security University of China, Beijing 100038, China)

Abstract: In recent years, deepfake has posed a severe challenge to the authenticity of online media and aroused public concern. Forged videos in real-world scenes are usually composed of double forgeries of audio and visual modes. However, mainstream deep forgery detection methods are realized by capturing artifacts in visual single mode, and different forgery methods will generate different artifacts, which leads to limited performance and poor generalization ability of such methods. At the same time, the work of detecting using multimodal information does not make full use of audio information to discover the forgery of visual modes. Therefore, a multi-stream fusion video deepfake detection method based on self-supervised learning (MFAVNet) is proposed, which uses the inconsistency between audio and visual modes for multi-mode forgery detection. Specifically, a feature extraction network with three parallel branches of video, audio and audio-visual is constructed, and the visual features and audio features of the lip region are extracted by using the self-supervised pre-trained Audio-Visual HuBERT(AV-HuBERT) model. In order to supplement the overall visual features of facial regions, a visual-audio feature extractor based on ResNet architecture is designed, and it is fine-tuned and optimized on forged data sets. After feature fusion, the final classification is achieved by multi-layer perceptron. Experimental results on DeepfakeTIMIT and FakeAVCeleb data sets show that the proposed method has good generalization adaptability when detecting videos generated by various forgery technologies, and the performance of MFAVNet in deepfake detection is better than many mainstream advanced methods.

Key words: deepfake;self-supervision;multimodal;feature fusion;visual;audio

0 引 言

随着深度伪造(Deepfake)技术的持续发展,特别是基于扩散模型(DM)等人工智能方法的应用,生成高度逼真难以分辨的图像、音频与视频内容已成为现实[1]。语音可以通过文本到语音合成(Text-

基金项目: 高等学校学科创新引智基地资助项目(B20087)。
作者简介: 田玉政(2001—),男,硕士研究生,主要研究方向:深度伪造检测。
通信作者: 蔡满春(1972—),男,博士,教授,博士生导师,主要研究方向:信息安全。Email: caimanchun@ppsuc.edu.cn。
收稿日期: 2025-06-09

to - Speech Synthesis, TTS) 和语音转换 (Voice Conversion, VC) 算法生成,视频可以通过面部交换技术篡改原始视频,也是当今最常见的视觉深度伪造内容,这些技术使得普通人也能以极低的成本和门槛制作出令人信服的虚假内容,从而放大了深度伪造技术的危害[2-3]。因此,开发有效的检测机制以应对并降低相关风险,已成为当务之急。

大多数深度伪造视频检测方法采用监督学习进行训练,通常依赖大量的人工标注数据,还需要专门用于深度伪造检测的标注数据集[4]。对标注数据的高度依赖限制了这类方法对网络上海量无标签视频资源的利用能力。此外,现有多数深度伪造检测方法仅关注单一模态,尤其侧重于分析视频中的视觉伪影,依赖深度学习模型对输入进行"真"或"假"的二元分类,缺乏对多模态伪造视频数据集的评估,较少考虑音频模态的融合[5]。然而,随着多模态深度伪造技术的发展,尤其是融合音频与视觉内容的生成方式的出现,伪造内容呈现出更高的真实感和多样性,特别是单一讲话者的视频,已成为虚假信息传播的重要载体。另一些研究是利用生物特征信号,例如识别特定个体特有的特定面部运动模式,但基于身份的检测方法在泛化至新身份时存在局限性[6]。

为实现更优的泛化检测能力,提高深度伪造检测精度并减少模型训练开销,本文提出了一种基于自监督学习的多流融合的视频伪造检测方法,在自监督预训练阶段不依赖复杂的视觉预处理,实验主要在多模态伪造数据集上进行。由于人脸置换或唇形同步等伪造操作引入的伪影,唇部运动与发音音节之间无法精确对齐,唇部生物特征和音频模态之间的不匹配可作为识别音视频深度伪造的重要线索。但仅基于唇部特征的伪造检测方法在唇部区域未被篡改或仅被轻微伪造的情况下存在鲁棒性不足的问题,为此本文还设计了一种基于 ResNet 的视频特征提取器,用于提取完整人脸的特征,从而为检测过程提供辅助信息;设计了具有视听、音频和视频三分支的多流框架;对预训练的自监督模型进行微调,从而有效地捕获音频和视觉模态之间的视听相关性与同步;在两个多模态深度伪造数据集上达到了最先进的性能水平,并且能够有效检测各种未知的伪造生成技术,展现出高度的泛化性。

1 相关工作

1.1 深度伪造检测方法

由于深度伪造技术本质是对真实人脸进行篡改伪造,因此在某些生物特征上通常与真实人脸存在差异,如眨眼频率、面部光线反射和嘴部运动等。针对眼部特征,Liy 等[7] 提出了一种结合卷积神经网络和循环神经网络的方法,利用眨眼频率异常进行检测,在自建数据集上 AUC 达到了 0.99;Jin 等[8] 指出在传统的光电容积脉搏波描记法(PPG) 检测伪造人脸视频中,为了消除环境变化造成的噪声、提取清晰的生理信号,通常会对视频进行去噪与滤波处理,但这会破坏伪造视频中的异常信号、造成低效性问题,因此提出将同一块中不同帧的信号值按行排列,再通过深度网络进行分类,达到基于不同心率提取算法的伪造人脸检测效果。由于这些方法通常针对单一生物特征,导致对不同类型伪造数据集的泛化能力较差。Knafo 等[9] 提出一种由自监督阶段和有监督微调阶段组成的新的多模态方法,在自监督阶段,利用域外多模态视频为每种模态创建鲁棒的表示;在有监督微调阶段,通过任务特定的检测器实现在伪视频检测方面的微调,该方法能够克服预训练阶段对大量未标记数据的需求,提高了模型在伪造视频检测任务上的性能;Haliassos 等[10] 提出基于视听觉的自监督学习检测方法,采用教师-学生模型思想,利用真实视频中视觉和听觉的自然对应关系,通过 BYOL(Bootstrap Your Own Latent) 框架学习时间密集的视频表示,以捕获面部运动和表情等特征。但基于多模态融合的检测模型可能会过拟合于训练数据中所包含的特定伪造生成方式,导致在面对真实场景中尚未见过的深度伪造算法时表现出较差的泛化能力。

1.2 多模态自监督学习

自监督学习(SSL) 在多模态学习领域取得了显著进展,尤其在无需人工标注数据的条件下实现了强大的表征能力。在音视频数据上进行的自监督预训练不仅提升了下游任务的性能,也为深度伪造检测提供了新的方向。Audio Visua- HuBERT(AV-HuBERT) 是一个基于 SSL 的视听表征学习模型,在唇读、视听语音增强、视听语音分离和视听语音识别多个任务上均取得了当前最优的性能[11]。受其在多个下游任务中最佳性能的驱动,本文引入 AV-HuBERT 进行特征提取,以捕捉感兴趣的唇部区域和对应音频信号之间的不一致性。

Alayrac 等[12] 提出的 MMV(MultiModal Versatile) 模型是另一种代表性的多模态自监督学习方法,旨在从未标注的视频中联合学习视觉、音频和文本模态的表征,通过对比学习机制将不同模态的信息映射到统

一嵌入空间中，从而实现模态间的相互对齐。MMV 的核心思想在于利用“音频-视觉-文本”三模态之间的天然同步性，挖掘跨模态之间的互补关系。尽管 MMV 并非专为深度伪造检测设计，但其在建模模态一致性方面的优势为伪造检测提供了启发。本文借鉴了 Alayrac 等的研究成果和提供的网络架构，实现深度伪造检测的目标，提升了模型的泛化能力。通过集成强大的视听表征、语音-嘴唇同步特征和时空面部特征，所提出的伪造检测方法在两个多模态深度伪造数据集上得到了更佳的性能。

2　本文方法设计与实现

在深度伪造检测的下游任务中，受限于多模态数据集数量较少以及正负样本分布不均等问题，现有检测方法常常面临跨伪造方法泛化性差、跨数据集泛化能力弱等挑战。而多模态自监督预训练模型在预训练阶段学习到每个模态的鲁棒表示，可通过少量标注数据进行微调而迁移至特定的下游任务中，以增强检测方法的泛化性和鲁棒性。基于此，本文提出一种基于自监督学习的多流融合的视频伪造检测方法（MFAVNet）。

2.1　总体框架结构

本文使用了一种在大规模域外视频数据集上以自监督方式预训练的多模态骨干 MMV 网络，应用于视频、音频特征提取网络上。采用预先训练的骨干，学习输出关于面部伪造域外数据每个模态的鲁棒表示，并使其适应深度伪造检测任务，通过对比学习方法学习特定模态的特征表示。视听特征提取器基于在 LRS3 数据集上预训练的 AV-HuBERT 模型，将在新的多模态深度伪造数据集上训练检测模型时进行微调，在微调过程中，骨干网络和分类器头部的权重均未冻结。

模型整体结构可以分为数据预处理、特征提取、特征融合和分类 4 个部分，如图 1 所示。在数据预处理模块中进行人脸提取和数据增强；特征提取模块包括 3 个特征提取器，即视频特征提取器、音频特征提取器和视听特征提取器，以捕获视觉和音频特征之间的时间相关性。特征融合模块用于整合 3 个分支产生的输出，使用包含两个隐藏层的多层感知机（Multilayer Perceptron，MLP）作为分类器。给定一个测试视频 x，最终融合特征会被输入至 MLP 中，根据下式判断视频真假：

$$\mathrm{MFAVNet} = D(C_{\mathrm{v}}(x), C_{\mathrm{a}}(x), C_{\mathrm{av}}(x)) \quad (1)$$

其中，C_{v}，C_{a} 和 C_{av} 分别表示纯视频、纯音频和视听特征提取器输出的特征向量表示，D 表示决策分类的功能。

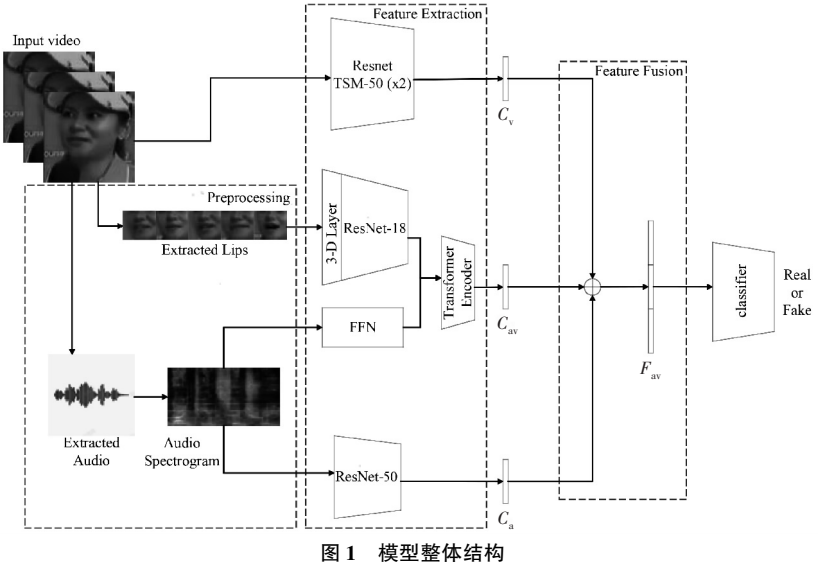

图 1　模型整体结构

Fig. 1　Overall model architecture

2.2　数据分析和预处理

本文方法主要利用人脸图像、唇部图像和音频信息进行多模态伪造检测。从每个视频中按 29 fps 的速率提取连续的 32 帧作为视频特征提取网络的输入；采用多任务级联卷积神经网络（MTCNN）对每帧图像进行初步的人脸检测，获得粗略的面部边界框；对检测到的边界框进行适当扩展，以覆盖更多面部区域并保留细节信息；在此基础上，利用 MediaPipe 开源多媒体数据处理框架提取面部关键点，并结合平均人脸形状对每一帧中的人脸进行对齐操作，为每个视频生成一个图像序列，包含对齐居中的 224×224 RGB 人脸图像帧。

从每个视频中提取采样率为 16 kHz 的音频轨道,从音频波形中提取对数梅尔频率倒谱图,用作音频网络的输入,这些梅尔频率倒谱图也被用作视听网络中的音频模态输入。对于基于 AV-HuBERT 的视听网络,使用面部标志点从正面人脸图像中提取相应的嘴唇区域,得到 96×96 的 RGB 嘴唇图像序列,并在输入网络前转换为灰度图像,形状为 $C \times F \times H \times W$,其中 C 表示通道数,F 表示帧数,H 和 W 分别表示每帧的高度和宽度,将图像特征序列与梅尔频率倒谱图特征配对,作为网络的输入。

在模型的微调过程中,为了提高模型的泛化能力和鲁棒性,对每个从视频中提取的剪辑应用数据增强方法。随机地对视频剪辑实施了水平翻转和颜色增强两种视觉增强策略,为了模拟真实世界中的听觉环境,向音频信号中添加噪声以进行音频增强,但在模型的推理阶段,没有应用任何数据增强的方法,以确保模型能够直接处理原始数据。

2.3 视频和音频特征提取网络

在给定一组多模态且未标注视频的条件下,MMV 旨在学习一个能够处理任一训练模态并输出可用于与其他模态进行比较的表征的模型。一个视频样本 x 包含 3 种模态组成的集合,记为 $x = \{x_v, x_a, x_t\}$,其中 x_v, x_a, x_t 分别表示视觉、音频与文本模态,这些模态分别对应于 RGB 帧序列、音频采样以及通过预训练自动语音识别(ASR)系统获得的离散词标记。针对每种模态,都有一个模态特定的主干神经网络 f_m,接收视频的某个模态 x_m 作为输入,并输出一个维度为 d_m 的表征向量。模态 m 的表征向量记作 $z_m = f_m(x_m)$,而所有模态的表征向量集合则记为 $z = f(x)$,其中 $z = \{z_v, z_a, z_t\}$。MMV 的目标是使不同模态间的表征易于进行相互比较,并用于损失函数的计算,因此各模态的表征将通过投影头被映射至一个共享空间 $S_s \subset R^{d_s}$ 中,其中 s 表示被嵌入该共享空间的模态集合,例如 $s = va$ 表示视觉和音频联合空间。不同模态的表征可以通过点积进行比较。输入模态 x_m 在共享空间 S_s 中的表示记作 $z_{m \to s}$,利用跨模态表征以获得先进的实验结果。

对于人脸图像序列,使用所有通道加倍后的 ResNet50(×2)网络的时间移位模块(Temporal Shift Module,TSM)作为视频特征提取器,在视频分支的最后一层应用时间和空间平均池化,获得单个视觉表示向量 $\boldsymbol{C}_v$,该过程可表示为:

$$\mathrm{GAP} = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} v_{h,w} \tag{2}$$

$$\boldsymbol{C}_v = \frac{1}{T} \sum_{t=1}^{T} \mathrm{GAP} \tag{3}$$

其中,H、W 分别为特征图的高度和宽度;GAP 为在空间上的全局平均池化的结果;T 为时间帧数。

采用 32 帧 $v \in R^{32 \times 3 \times 224 \times 224}$,224 表示每个视频帧的高度和宽度,3 表示 RGB 颜色通道,为基于 ResNet TSM-50(×2)的骨干创建维度为 4 096 的视觉表示向量。

对于音频模态,首先从输入视频中提取 WAV 格式的音频。WAV 文件是具有原始格式的音频,不能直接传递到检测模型中,因此原始音频波形被转换成具有 80 个频率区间的对数梅尔频率倒谱图,其中时间和频率元素有助于学习和捕获声学模式、时间动态和其他音频特定特征,并在音频分支的最后一层应用空间池化以获得音频表示向量 $\boldsymbol{C}_a$:

$$\boldsymbol{C}_a = \frac{1}{T' \times F'} \sum_{t=1}^{T'} \sum_{f=1}^{F'} a_{t,f} \tag{4}$$

其中,T' 为时间帧总数,F' 为频带数量。

根据音频信号采用 ResNet-50 模型处理 MFCC 特征,并且和视频帧是同步进行的,音频模态向量的维度为 2 048。

2.4 视听特征提取网络(AVN)

利用视觉和音频模态,通过联合学习成对的音频和唇部图像序列输入来提供补充信息,并捕捉不同模态之间有意义的表示,以揭露视频中的伪造。视听特征提取器由一个 Resnet-18、一个轻量级的前馈网络(FFN)和一个 Transformer 编码器组成。

Resnet-18 网络用于从连续输入的嘴唇图像帧中提取基于唇部的视觉特征,其过程可以表示为:

$$F_v = \mathrm{ResNet18}(L) \tag{5}$$

其中,L 表示输入的唇部图像帧序列。

FFN 用于从音频梅尔频率倒谱图中提取音频特征,其过程可以表示为:

$$F_a = \mathrm{FFN}(M) \tag{6}$$

其中,M 为梅尔频率倒谱图。

视觉和音频特征沿着特征维度融合,并馈送到共享的 Transformer 编码器,以提取封装了音频和视觉模态之间相关性的上下文视听表示,该编码器通过下式生成维度为 1 024 的视听嵌入向量 $\boldsymbol{C}_{av}$。

$$\boldsymbol{C}_{av} = \mathrm{AVN}(F_v, F_a) \tag{7}$$

其中,F_v 代表帧级别的视觉特征,F_a 代表对应的音频特征。

2.5 基于 MLP 的分类

本文采用晚期融合中特征融合的方法,通过模

型分别提取各自模态的特征，各模态相互独立，特征处理和学习过程互不干扰，生成特征向量，最后将其结果融合。特征融合方式有 2 种，一种是两个有相同形状的特征向量在元素级别上对应相加，简称 add；另一种是把特征图堆到一起，简称 Concat。本文特征融合方式基于 Concat 形式，音频特征、视觉特征、视听特征将在特征维度上进行拼接，如下式，组合成了一个更宽的特征向量 $\boldsymbol{F}_{av}$，总维度为 7 168。

$$\boldsymbol{F}_{av} = \boldsymbol{C}_{v} \oplus \boldsymbol{C}_{av} \oplus \boldsymbol{C}_{a} \tag{8}$$

其中，$\oplus$ 代表拼接操作。

使用一个 3 层的多层感知机作为二分类器。为更好利用模型主干提取到的多模态特征，直接将模型主干的融合特征 $\boldsymbol{F}_{av} \in R^{1\times 7168}$ 作为分类层的输入。多层感知机由输入层、两个隐藏层和输出层组成，本文 MLP 的两个隐藏层分别设置为 512 和 128，输出结果为视频帧为假的概率。分类层的计算过程可表示为：

$$y = \sigma(W_3(W_2(W_1\boldsymbol{F}_{av} + b_1) + b_2) + b_3) \tag{9}$$

其中，σ 为 Softmax 函数。

3　实验与分析

在两个数据集 FakeAVCeleb 和 DeepfakeTIMIT（DF-TIMIT）上进行了实验。与其他单模态伪造数据集不同，这两个数据集同时具有音频和视觉模态，假样本中包含音频或视觉伪造。此外，两个数据集中的视频中的人脸是正面的，有利于稳定地提取唇部区域帧。

3.1　实验环境

本文实验的环境配置信息见表 1。

表 1　实验环境

Table 1　Experimental environment

类别	配置
CPU	12 vCPU Intel(R) Xeon(R) Platinum 8352V CPU @ 2.10 GHz
GPU	vGPU-32 GB(32 GB)
操作系统	ubuntu20.04
编程语言	Python
CUDA	11.8

3.2　实验参数

本文的损失函数为二元交叉熵损失函数，其定义如下：

$$\boldsymbol{L} = y\log\hat{y} + (1-y)\log(1-\hat{y}) \tag{10}$$

在实验设置中，模型采用了 Adam 优化器并结合了余弦衰减学习率调度策略，进行了 30 个 epoch 的训练。初始学习率设定为 1e-5，衰减参数固定为 0.95。

本文使用准确率（Accuracy，ACC）和 AUC（Area Under the Curve）作为评判指标，来评估单个单模态/多模态分类器和本文提出的多模态 MFAVNet 分类器的性能。对于所有指标，值越高表示性能越好。准确率的定义：

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FN} + \text{FP} + \text{TN})} \tag{11}$$

3.3　数据集

FakeAVCeleb 数据集是于 2022 年发布的视听数据集，专门用于深度伪造检测任务，基于 4 种最新的深度伪造和合成语音生成技术，包括面部交换（Faceswap）、Fsgan（Forensic Style GAN）、语音驱动口型同步模型（Wav2Lip）和实时语音克隆（Rtvc）。该数据集包括 4 个类别：RARV（真音频、真视频）、FARV（假音频、真视频）、RAFV（真音频、假视频）和 FAFV（假音频、假视频）。由于数据集官方未提供正式的数据集分割方法，以前的工作通常通过主题 ID 或随机分割 FakeAVCeleb 数据集，但这些分割方法在评估模型对于未知的深度伪造生成方法的泛化性能力具有局限性。本文提出了一种新的分割机制：基于伪造生成方法来分割数据集，以评估不可见伪造生成方法的性能。在创建训练集、验证集和测试集的过程中，其比例为 7∶1∶2，确保从训练集和验证集中排除测试集中使用的生成方法。FakeAVCeleb 数据集示例如图 2 所示。

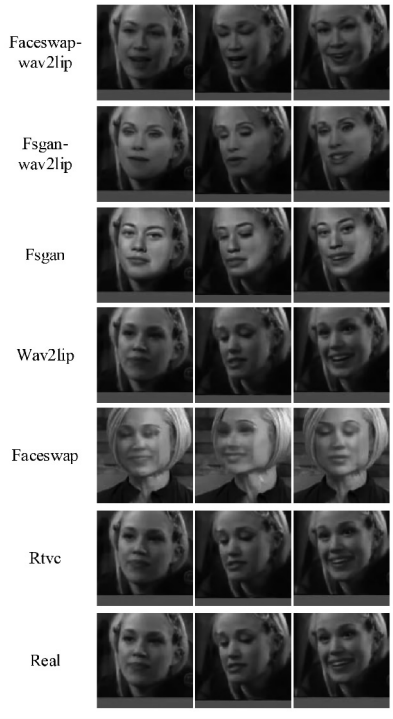

图 2　FakeAVCeleb 数据集示例

Fig. 2　FakeAVCeleb dataset example

训练集包含 350 个视频,分别来自类别 RARV、FARV、RAFV 和 FAFV(不包括 faceswap 和 faceswap-wav2lip)。验证集包含 50 个视频,分别来自 RARV,FARV,RAFV 和 FAFV 类别(不包括 faceswap 和 faceswap-wav2lip)。对于测试集,采样了 100 个不包括在训练集和验证集中的 faceswap(RAFV)和 faceswap-wav2lip(FAFV)视频。为了防止信息泄露,实验的微调阶段中,训练集、验证集、测试集均不包含来自相同身份的视频。

DeepfakeTIMIT 是一个视听数据集,没有对音轨信息进行伪造。该数据集包括 640 个视频,于 2018 年公开,包含两个子集,即 64×64 分辨率大小的 LQ(低质量)集以及 128×128 大小的 HQ(高质量)集。在测试过程中均采用 HQ 集,从 VidTIMIT 数据集上下载了真实视频进行实验,确保正负样本平衡,真视频和假视频中的音频都是真实的。两个多模态数据集的信息见表 2。

表 2 多模态数据集信息

Table 2 Multimodal dataset information

数据集	真实/伪造数量	伪造生成方法	是否包含音频	公开可用性
FakeAVCeleb[13]	500/20 000	Faceswap, Fsgan, wav2lip, RTVC	是	是
DeepfakeTIMIT[14]	0/640	Faceswap	是	是

3.4 结果分析

单模态检测模型是仅使用视觉特征进行训练用于检测视频帧之间的伪影,对比的模型有 Xception[15]、CViT[16]、Lip Forensics[17] 和 Face X-ray[18]。多模态检测模型使用音频特征(如 MFCC)和视觉特征联合学习进行伪造检测,对比的模型包括 VFD[19]、AVA-CL[20]、AVT^2-DWF[21]、AVOiD-DF[22]、MRDF-Margin[23] 和 EmoForen[24]。

3.4.1 域内实验结果

在测试实验中,对照最先进的基线方法来评估 MFAVNet 方法的性能,基线方法分为 2 组:视觉模态(V)和多模态(AV)。在 FakeAVceleb 多模态伪造数据集上进行微调,多模态伪造数据集上的结果见表 3,MFAVNet 方法准确率达到了 93.20%,AUC 达到了 94.80%,高于所有基于视觉单模态的方法,比使用音频信息中表现最好的方法 MRDF-Margin 的 AUC 提升了 3 个百分点,验证了该方法的可行性。在该数据集上的分类结果均优于其他单模态或多模态的检测方法,表明本文所提出的方法对音频信息利用更加充分,同时融合人脸和唇部视觉特征,提升了模型对跨模态伪造痕迹的感知能力,具备更好的检测能力。

表 3 对比实验结果

Table 3 Comparison experiment results

方法	模态	FakeAVceleb 上 ACC/%	FakeAVceleb 上 AUC/%
Xception	V	72.71	73.51
CViT	V	75.14	79.00
Lip Forensics	V	80.10	82.40
Face X-ray	V	72.88	73.52
VFD	AV	81.52	86.11
AVA-CL	AV	86.55	89.47
AVT^2-DWF	AV	87.57	88.32
AVOiD-DF	AV	83.70	89.20
MRDF-Margin	AV	93.40	91.80
MFAVNet	AV	93.20	94.80

3.4.2 跨域实验结果

视频通常不严格服从训练集分布,检测模型需要应对包括伪造类型、光照变化等多种域间差异。为了验证模型的泛化能力,在 FakeAVCeleb 训练集上对模型进行了微调,并在 DF-TIMIT 数据集上进行了性能评估。DF-TIMIT 数据集发布时间较早,伪造方法造成的伪造痕迹较重,因此所有方法的 AUC 均达到了不错的数值。跨数据集实验结果见表 4,MFAVNet 准确率达到了 99.90%,AUC 达到了 99.90%,在性能上达到了先进水平,均优于其他对

比方法，表明采用自监督预训练的模型，在预训练阶段从大规模未标注数据中学习到通用的特征表示，迁移到深度伪造检测下游任务上，具有较强的跨数据集泛化性。

表 4　跨数据集实验结果

Table 4　Cross-dataset experiment results

方法	模态	DF-TIMIT 上 ACC/%	DF-TIMIT 上 AUC/%
Xception	V	95.20	95.60
CViT	V	98.01	98.73
Lip Forensics	V	99.25	99.27
Face X-ray	V		94.47
EmoForen	AV		94.90
VFD	AV		99.82
AVA-CL	AV	96.53	99.86
AVT^2-DWF	AV	98.43	98.43
MFAVNet	AV	99.90	99.90

3.4.3　消融实验

在对 MFAVNet 模型进行全面的性能评估时，实施了一系列消融实验。实验涵盖了纯视觉模态、AV 模态即通过简单的将音频与人脸面部特征相结合以及集成了 AV 和 AVN 模块的 MFAVNet 版本。在 FakeAVCeleb 数据集上的消融实验结果见表 5。

表 5　消融实验

Table 5　Ablation experiments

视频模态	音频模态	AVN	FakeAVCeleb 上 AUC/%
启用			91.20
启用	启用		91.70
启用	启用	启用	94.80

在 FakeAVCeleb 数据集中，部分视频的视觉模态为真实，而音频模态经过伪造。实验结果表明，音频和视觉模态的特征融合显著提升了检测性能。当引入 AVN 模块后，可以更精细地捕捉伪造的局部纹理与动态异常，检测结果的 AUC 分数提升了 3.6%，充分证明了 MFAVNet 方法在提取不同模态共享特征方面的显著优势。

4　结束语

本文从多模态视听的角度研究了视频伪造检测的下游任务。使用了视觉和音频两种模态，以及对特定下游任务不变的自监督骨干网络，设计了一种自监督学习的深度伪造检测方法 MFAVNet。本文所提的方法联合利用 SSL 音频/视觉/视听表示以及嘴唇图像帧和音频之间的时间相关性来检测深度造假内容，并将其与各种现有的单模态和多模态模型进行了比较，在 FakeAVCeleb 和 DeepfakeTIMIT 数据集上都取得了前沿的性能。未来的工作将致力于探索结合 Transformer 模型的深度伪造检测方法，探究视频帧局部特征和全局特征的融合以及多模态学习方式，提高在深度伪造检测领域的检测性能和准确性。

参考文献

[1] CROITORU F A, HONDRU V, IONESCU R T, et al. Diffusion models in vision: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 10850-10869.

[2] TAN X, QIN T, SOONG F, et al. A survey on neural speech synthesis[J]. arXiv preprint arXiv,2106.15561, 2021.

[3] SISMAN B, YAMAGISHI J, KING S, et al. An overview of voice conversion and its challenges: From statistical modeling to deep learning[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 29: 132-157.

[4] AGARWAL S, FARID H, FRIED O, et al. Detecting deep-fake videos from phoneme-viseme mismatches[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2020: 660-661.

[5] WANG S, WANG O, ZHANG R, et al. CNN-generated images are surprisingly easy to spot... for now[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 8695-8704.

[6] AGARWAL S, FARID H, GU Y, et al. Protecting world leaders against deep fakes[C]// Proceedings of CVPR Workshops. Piscataway, NJ: IEEE, 2019, 1(38):90-98.

[7] LIY C M, INICTUOCULI L. Exposing ai created fake videos by detecting eye blinking [C]//Proceedings of 2018 IEEE International Workshop on Information Forensics and Security (WIFS). Piscataway, NJ: IEEE, 2018:1-6.

[8] JIN Xinlei, YE Dengpan, CHEN Chuanxi. Countering spoof: Towards detecting deepfake with multidimensional biological signals[J]. Security and Communication Networks, 2021, 2021(1): 1-10.

[9] KNAFO G. Fakeout: Leveraging out-of-domain self-supervision for multi-modal video deepfake detection[D]. Israel: Reichman University, 2022.

[10] HALIASSOS A, MIRA R, PETRIDIS S, et al. Leveraging real talking faces via self-supervision for robust forgery detection [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 14950-14962.

[11] SHI B, HSU W N, LAKHOTIA K, et al. Learning audio-visual speech representation by masked multimodal cluster prediction[J]. arXiv preprint arXiv, 2201.02184, 2022.

[12] ALAYRAC J B, RECASENS A, SCHNEIDER R, et al. Self-supervised multimodal versatile networks[J]. Advances in Neural Information Processing Systems, 2020, 33: 25-37.

[13] KHALID H, TARIQ S, KIM M, et al. FakeAVCeleb: A novel audio-video multimodal deepfake dataset [J]. arXiv preprint arXiv, 2108.05080, 2021.

[14] KORSHUNOV P, MARCEL S. Deepfakes: A new threat to face recognition? assessment and detection[J]. arXiv preprint arXiv: 1812.08685, 2018.

[15] ROSSLER A, COZZOLINO D, VERDOLIVA L, et al. Faceforensics++: Learning to detect manipulated facial images [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2019: 1-11.

[16] WODAJO D, ATNAFU S. Deepfake video detection using convolutional vision transformer[J]. arXiv preprint arXiv, 2102.11126, 2021.

[17] HALIASSOS A, VOUGIOUKAS K, PETRIDIS S, et al. Lips don´t lie: A generalisable and robust approach to face forgery detection [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 5039-5049.

[18] LI Lingzhi, BAO Jianmin, ZHANG Ting, et al. Face x-ray for more general face forgery detection [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 5001-5010.

[19] CHENG H, GUO Yangyang, WANG Tianyi, et al. Voice-face homogeneity tells deepfake[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 20(3): 1-22.

[20] ZHANG Yibo, LIN Weiguo, XU Junfeng. Joint audio-visual attention with contrastive learning for more general deepfake detection [J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2024, 20(5): 1-23.

[21] WANG Rui, YE Dengpan, TANG Long, et al. AVT^2-DWF: Improving deepfake detection with audio-visual fusion and dynamic weighting strategies[J]. IEEE Signal Processing Letters, 2024,31:1-6.

[22] YANG Wenyuan, ZHOU Xiaoyu, CHEN Zhikai, et al. Avoid-df: Audio-visual joint learning for detecting deepfake[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 2015-2029.

[23] ZOU Heqing, SHEN Meng, HU Yuchen, et al. Cross-modality and within-modality regularization for audio-visual deepfake detection [C]// Proceedings of ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2024: 4900-4904.

[24] MITTAL T, BHATTACHARYA U, CHANDRA R, et al. Emotions don´t lie: An audio-visual deepfake detection method using affective cues [C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 2823-2832.