基于说话行为相关面部关键点的鲁棒伪造人脸检测方案

黄逸焕[1], 彭 荔[1], 任延珍[1,2], 王丽娜[1,2]

[1]武汉大学国家网络安全学院 武汉 中国 430072
[2]空天信息安全与可信计算教育部重点实验室 武汉 中国 430072

摘要 生成式音视频的日趋逼真给伪造检测带来了巨大挑战, 从社交媒体平台的虚假视频发布, 到政治宣传中的误导性内容, 其潜在风险无处不在。因此, 针对伪造说话人脸视频的有效检测与防范机制变得尤为迫切和重要。然而, 现有的主流深度伪造检测方法存在无法区分压缩痕迹和伪造痕迹的问题, 导致其在检测高度压缩的视频和社交通信场景时准确率显著下降。为了解决这一问题, 本文提出了一种基于说话行为相关面部关键点的鲁棒伪造说话人脸检测方案 FALNet(Facial-Landmark based Graph Attention Network)。本文通过分析说话过程中的肌肉动态, 设计了一个基于面部肌肉运动的邻接矩阵。该矩阵不仅保留了面部的拓扑信息, 而且可以有效捕捉真假面部特征之间的差异。另外, 考虑到时间特征在视频伪造检测中的重要性, 本文同时对长短时特征进行建模。具体来说, FALNet 首先使用图注意力网络捕获短时特征, 随后将短时特征序列输入循环神经网络以进行长时特征建模。实验结果表明, FALNet 在多个公开数据集的视频伪造子集上均取得了超过 98%的检测准确率, 相比于现有基于面部关键点的先进方法, FALNet 的 AUC 值(Area Under the Curve)取得了 0.6%~1.1%的提升, 且在面对压缩时检测的 AUC 值保持在 94%以上, 具有良好的鲁棒性。

关键词 伪造说话人脸检测; 深度学习; 鲁棒性; 面部关键点
中图法分类号 TP309 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2026.03.06

A Robust Forged Face Detection Scheme Based on Speech-Related Facial Landmarks

HUANG Yihuan[1], PENG Li[1], REN Yanzhen[1,2], WANG Lina[1,2]

[1] School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China
[2] Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, Wuhan 430072, China

Abstract The generation of synthetic audio-visual content is becoming increasingly realistic, posing significant challenges to the detection of falsified video. From the dissemination of fake audiovisuals on social media platforms to misleading content in political propaganda, the potential risks are pervasive. Consequently, the need for effective detection and prevention mechanisms against forged speaker facial videos has become urgent and crucial. However, current mainstream deepfake detection methods struggle to differentiate between compression artifacts and forgery artifacts, leading to a significant drop in detection accuracy in scenarios involving highly compressed videos and social media communications. We propose a Facial-Landmark based Graph Attention Network (FALNet) for detecting forged speaker facial videos, which decouples facial landmarks from video. We introduce a robust video feature extraction network based on facial landmarks and analyze the muscle movements associated with speech behavior, as well as the forged cues introduced during the generation of deepfake speaker facial videos. We designed an adjacency matrix based on facial muscle movements by analyzing the muscle dynamics during speech. The matrix not only preserves the topological information of the face but also effectively captures the differences between genuine and fake facial features. Using a graph attention network as the backbone, we extracted facial features represented by this adjacency matrix. Furthermore, considering the importance of temporal features in video forgery detection, we modeled both short-term and long-term features. Specifically, we first used a graph attention network to capture short-term features and then fed the sequence of short-term features into a recurrent neural network to model long-term dependencies. Experimental results show that our scheme has achieved a detection accuracy of over 98% on video forgery subsets of multiple public datasets. Compared to existing advanced methods based on facial key points, this scheme has achieved a 0.6% to 1.1% improvement in AUC (Area Under the Curve) value. Furthermore, when facing compression, the detection AUC value of this scheme remains

通讯作者: 任延珍, 武汉大学国家网络安全学院, 教授, Email: renyz@whu.edu.cn。
本文受到国家自然科学基金(No. 62172306, No. 62372334) 以及湖北省重点研发计划(No. 2021BAB018)支持。
收稿日期: 2024-07-19; 修改日期: 2024-12-11; 定稿日期: 2026-01-26

above 94%, demonstrating good robustness.

Key words deepfake detection; deep learning; robustness; facial landmarks

1 引言

近年来, 深度伪造技术(DeepFake)的迅速兴起, 给政治安全、经济安全、国民安全、社会安全等重大安全领域带来了诸多风险。随着技术的不断进步与普及, 一些不法分子开始利用深度伪造技术制造虚假内容, 企图达到欺诈、散播谣言、诽谤他人等目的。作为深度伪造技术的一个重要分支, 说话人脸伪造技术, 通过生成逼真的说话人虚假视频使得公众难以辨真假, 在数字化时代的信息传播中, 达到扭曲个人形象、误导公众信息的恶意目的, 对社会稳定和公共安全构成威胁。目前, 说话人脸伪造技术已开始渗透到在线实时交互场景中, 例如各种主流社交通信应用中的语音通话、视频通话、视频会议等。这种渗透性意味着攻击者可以利用深度伪造技术获得通话方对其身份的完全信任, 从而实施巨额的诈骗行为。说话人脸伪造技术不仅对个人隐私构成严重威胁, 还给商业和政府等机构带来了新的安全挑战。研究表明, 通过深度伪造技术生成的伪造视频很难被人们察觉和识别[1]。尤其在在线交互场景中, 人们更难在很短时间内辨别出深度伪造视频。因此, 迫切需要采取措施来防范和应对说话人脸伪造技术可能带来的风险。

为了抵御深度伪造技术带来的潜在危害, 针对伪造说话人脸的检测技术获得了广泛的关注, 基于音、视频的深度伪造检测技术在此领域展现出了显著的效果和潜力。然而, 现有伪造说话人脸检测技术存在鲁棒性较差的问题, 制约了伪造检测技术的实用化进程。这是因为现有主流的伪造说话人脸检测技术是基于像素级的特征。在在线交互场景中, 为了实现快速网络传输以保证通信双方良好的实时视听觉交互感受, 视频数据通常需要经过压缩编码, 接收方的视频会有一定程度的降质, 压缩后的深度伪造视频中存在与篡改伪影相似的压缩伪影。而像素级特征的伪造检测技术无法准确区分篡改伪影和压缩伪影, 导致这些方案在图像压缩、裁剪、噪声等常见的视频处理场景下变得十分脆弱。实验结果表明基于像素级特征的检测方案, 在视频重压缩场景下的鲁棒性较差, 伪造视频检测精度的下降幅度可达10%。因此, 探索稳定高精度的鲁棒检测算法显得尤为重要。

本文针对当前人脸深度伪造检测在压缩场景下不鲁棒的问题, 提出了一种基于说话行为相关面部关键点的鲁棒伪造说话人脸检测方案 FALNet (Facial-Landmark based Graph Attention Network), 本文的主要贡献包括: 1) 针对压缩导致检测精度大幅下降的问题, 本文基于说话行为所带来的肌肉运动以及深度伪造说话人脸视频生成过程带来的伪造线索, 构建了一个新的说话行为相关的面部关键点连接网络结构, 以图注意力网络为主干网络, 在保留面部拓扑结构的基础上, 实现了对面部真伪特征的提取; 2) 针对面部关键点特征表示的鲁棒性, 本文提出了基于时空图注意力机制的面部关键点特征表示网络, 引入了图注意力网络和门控循环单元分别实现对短时和长时伪造面部痕迹的捕捉。综合实验结果表明, 本方案在多个公开数据集的视频伪造子集上均取得了更优的检测性能, 具有良好的鲁棒性。

2 相关工作

基于视觉特征的深度伪造检测方案通常采用两类关键特征: 帧内特征和帧间特征。帧内特征主要侧重于分析单个视频帧内的像素级细节, 包括检测图像中的瑕疵、不自然的纹理、光线变化等。通过对比原始视频和伪造视频之间的差异, 从而寻找出伪造痕迹。而帧间特征则更注重于分析视频序列中不同帧之间的连续性和一致性。通过综合考虑这两种特征, 深度伪造检测系统能够更全面地识别出可能存在的伪造内容。

2.1 基于帧内特征的深度伪造检测方案

帧内特征常常使用空间特征进行检测。这种方法将视频分解为帧, 对每一帧图像进行检测, 检测信号包括空域和频域等方面。由于伪造视频往往会破坏原始图像在空间域和频率域的自然分布特性, 因此这类方法具有一定的可靠性。

基于空域特征的检测方案重点关注生成的人脸在眼睛、鼻子等部位可能存在的颜色不一致和光照不一致性, 例如, 生成的视频中可能会缺少眼睛的反射细节, 牙齿细节的位置也可能存在缺失。同时, 人脸区域与周围背景的分辨率也可能表现出不一致。基于深度神经网络的早期工作使用 CNN, XceptionNet 等网络提取像素级的特征[2-3]。文献[4]通过结合注意力层和孪生范式, 利用集成的 CNN 模型, 实现了人脸伪造检测。可视化的结果表明网络的注意力集中于眼睛、嘴巴和鼻子等区域。文献[5]中提

出两个 Convolutional-Transformer 混合结构的模型, 将 EfficientNet B0 作为基于卷积的特征提取器, 结合 ViT 和 CrossViT, 在时间上和跨多个人脸上进行投票判决, 推断出视频片段的真伪。这项工作中, 将卷积网络作为图像特征的前端提取器, 通过卷积网络从图像中提取重要和低级的局部信息, 从而简化了后端 ViT 的训练。考虑到伪造痕迹可能出现在图像的全局也有可能是图像的局部, 将图像分成固定大小的小块可能损失了一部分全局特征, 因此在 CrossViT 方案中, 使用多尺度 Transformer 结构, 获取更加完整的全局和局部信息。该方案在伪造数据集上取得了优秀的结果, 表明空域特征能实现简单有效的检测。

目前, 由于基于生成对抗网络(Generative Adversarial Network, GAN)的生成模型在伪造过程中会经历上采样的过程, 因此伪造图像的频域与真实图像的频域存在明显的差异[6], 部分工作提出了在图像频域寻找伪造痕迹的方案。基于频率的方法通过离散余弦变换(Discrete Cosine Transform, DCT)将输入转换为频域来分析输入差异[7-9]。Liu 等人[10]提出了 SPSL 模型, 利用上采样过程中频域的变化, 将原始图像和相位谱作为输入实现对人脸伪造的检测, 以提高人脸伪造检测的泛化能力。

2.2 基于帧间特征的深度伪造检测方案

近年来, 除了空间伪影外, 时间伪影也成为研究人员关注的焦点。由于人脸视频的深度伪造过程, 会将原始视频划分为多个帧, 并对每一帧进行独立伪造, 这种处理方式缺乏对上下文的考虑, 可能导致生成视频帧在时间和空间上的不连续, 例如眨眼[11]、嘴唇运动[12]和面部关键点[13]的不连续。这些不连续性被用作深度伪造检测的有效线索。而帧间特征的加入综合了空间与时间两个维度的不一致性。通过提取帧间特征, 弥补仅利用空间特征所带来的时序特征缺失的问题, 使特征提取更加丰富。文献[14]提出了一种基于双分支网络结构的深度伪造检测方法, 该方法通过学习放大伪影, 同时抑制高级人脸内容, 实现伪造人脸的检测。文献[15]利用时序的视觉 Transformer 模型, 来探索视频片段中的长时帧间的不一致性。

视频压缩会导致深度伪造视频出现类似于篡改伪影的压缩伪影。因此, 从压缩的伪造视频中提取出具有辨识度的特征变得更加困难。目前, 大多数深度伪造检测方案都依赖于像素级的特征。然而, 在压缩等常见的视频处理情况下, 这些方案的检测精度可能会显著下降。由于面部关键点只记录不同面部关键点的坐标, 而关键点不受外部扰动的影响, 因此对传输过程中的各种扰动更鲁棒。已经有工作关注面部关键点以及面部关键点的动态特征[16-18], 用于实现一个更加鲁棒的深度伪造视频的检测。

文献[13]提出了一个高效且鲁棒的框架, 通过捕捉几何特征并进行时间建模, 来检测伪造人脸视频。此工作首先对视频的每一帧进行人脸检测, 并保留人脸的重点区域。在裁剪人脸图像后, 使用 OpenFace[19]提取 68 个面部关键点, 并进行关键点矫正, 勾勒出人脸的标志性轮廓。对检测到的 68 个面部关键点坐标, 分别构建成两个特征向量。即将全部帧的坐标作为输入特征, 以及将相邻帧坐标变化的差值作为输入特征。采用两个循环神经网络(Recurrent Neural Network, RNN)分别处理这两个特征向量, 在最后使用全连接层进行各自的分类, 最终, 取两者平均值作为分类结果。第一个 RNN 对面部形状的移动模式进行建模, 模拟面部形状运动特征; 第二个 RNN 对面部关键点差异进行建模, 用于捕捉时间的不连续性。该方案在检测高度压缩或噪声损坏的视频方面有很好的鲁棒性。在面对高度压缩的视频时, 性能仅轻微下降。

像素级的视频特征在压缩场景下鲁棒性较差。虽然已经有工作对人脸的鲁棒特征如面部关键点等进行了研究, 但是研究并不深入, 对面部关键点的选择和特征建模方式上仍然有研究和探索的空间。本文针对上述问题提出了针对性的解决方法。

3 基于说话相关面部关键点的鲁棒伪造人脸检测方案

本章对基于说话相关面部关键点的鲁棒伪造人脸检测方案的研究动机、算法总框架及相关技术细节进行介绍。

3.1 研究动机和算法总框架

在深度伪造视频检测中, 有损压缩可能会引起两方面问题。首先, 它可能导致视频中的大量信息丢失。这可能包括细微的特征和痕迹, 这些特征可能是检测深度伪造的关键线索。其次, 压缩过程可能会模糊、混淆或掩盖操纵的痕迹, 这使得检测深度伪造变得更加困难。面部关键点只记录坐标, 而关键点坐标不受外部扰动的影响, 因此对传输过程中的各种扰动更鲁棒。目前的研究工作已经证明了面部关键点特征在检测合成人脸图像或视频方面的潜力。然而, 现有的基于面部关键点的人脸深度伪造检测工作存在以下问题: 1) 基于面部关键点的特征提取在特征选择和建模方面缺乏深入研究。尽管已有研究引入面部关键点来提高检测鲁棒性, 但现有深度学习方

案存在局限性。比如, LRNet[13]使用循环神经网络输入每帧图像中的 68 个关键点坐标, 但未考虑关键点的空间结构, 忽略了人脸的拓扑信息。现有的伪造检测方法[16]使用图神经网络, 但未充分分析关键点之间的联系, 导致特征向量信息不足且计算量较大; 2)面部动作是连续的, 但当前工作对面部关键点的短时和长时特征缺乏充分提取和分析。短时特征捕捉局部表情变化, 如微表情, 时间范围为几百毫秒到数秒。长时特征则侧重于面部表情的整体变化, 涵盖几秒以上的时间。结合短时和长时特征分析, 可提高深度伪造检测的准确性和鲁棒性。

本研究针对当前人脸深度伪造检测方法在压缩场景下鲁棒性差的问题, 提出了一种基于面部关键点的人脸深度伪造检测网络 FALNet(Facial-Landmark based Graph Attention Network), 基于受压缩影响的较小的面部关键点特征, 构建了鲁棒的视频特征提取方案, 其整体结构如图 1 所示。首先, 本文通过对面部关键点的深入研究, 结合关键点本身的特质, 构建了一个基于说话相关关键点的面部图网络结构, 具体内容见 3.2 节; 针对视频中的短时和长时特征, 本文采用了时空图神经网络结合循环神经网络的方案, 以实现对视频人脸关键点特征的精准提取, 具体内容见 3.3 节。这种综合利用时空信息和神经网络技术的方法, 使得人脸深度伪造检测方法在复杂压缩场景下表现出更高的鲁棒性。

3.2 基于说话相关关键点的面部图网络结构

本文的面部关键点图网络构建方案, 主要从说话人脸的肌肉运动和深度伪造视频生成过程中带来的伪造痕迹, 来完成面部关键点网络的构建。相比于全连接的结构, 面部关键点网络的构建在减少计算量的同时, 保留面部的拓扑结构, 实现在关键点间更好的传递和聚合信息。

3.2.1 基于肌肉运动的面部关键点网络构建方案

不同的情绪和发音会导致不同的肌肉运动。如图 2 所示, 人在表达愤怒情绪时, 常常眉毛聚合下沉, 伴随瞪眼凝视。弯起的眉毛可能表示惊讶或不满, 而眼睛的眨动可能表示兴奋或焦虑。发音时, 嘴巴的形状会随着不同的音素而变化。例如, 发出“p”、“b”等闭合音时, 嘴唇会闭合, 发出“f”、“v”等摩擦音。

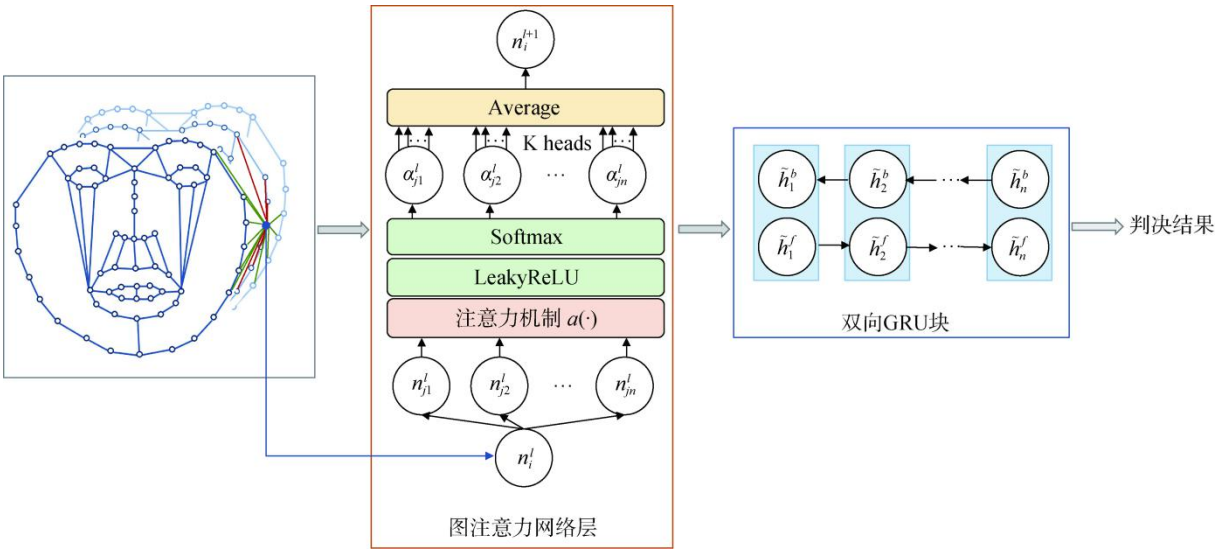
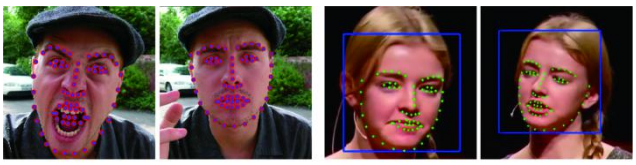


图 1 基于时空图注意力机制的面部关键点特征表示网络结构图

Figure 1 Network architecture diagram of facial landmark feature representation based on spatiotemporal graph attention



(a) 情绪相关肌肉运动 (b) 发音相关肌肉运动

图 2 面部肌肉运动示意图

Figure 2 Diagram of facial muscle movements

在生理学和心理学的研究中, 面部肌肉运动被认为由一组离散的组件组成, 这些组件被称为动作单元(Action Units, AU)。参考面部表情识别相关工作[20], 本节将额部(AU4、AU9)、眼睛(AU1、AU2、AU5、AU6)、鼻子(AU9)和嘴部区域(AU12-AU28)的关键点及连接添加进面部关键点网络中。关键点之间的连接是根据它们的语义关系定义的。这些重点区域代表了人脸发音和情绪相关的关键肌肉, 能

够捕捉到不自然的面部肌肉活动, 从而为人脸伪造检测提供线索。

根据上述思路, 本文构建出的面部肌肉运动网络如图 3 所示。具体而言, 本文使用 OpenFace 提取的 68 个面部关键点(LM_i, i=1-68)的位置信息, 使用的关键点包括额部关键点(LM_i, i =17-21, 22-26)、眼睛关键点(LM_i, i =36-41, 42-47)、鼻子区域(LM_i, i =27-35)和嘴部区域地标(LM_i, i =48-59, 60-67)。将这些区域连接起来, 代表五官以及面部轮廓的运动形态。这样的面部网络构建方案充分利用了面部区域的语义信息, 保留了面部的拓扑结构, 能够更准确地建模关键的面部动作单元以及它们之间的关系, 从而为捕捉人脸发音时的五官不协调和不自然的微表情提供更可靠的基础。

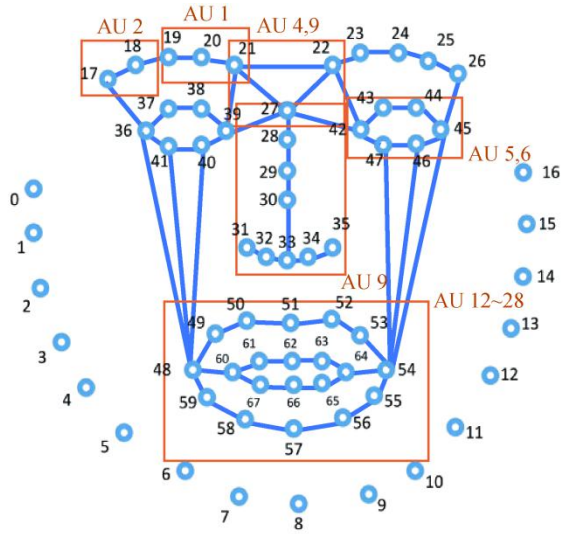


图 3 基于肌肉运动面部关键点网络结构设计

Figure 3 Design of facial landmark network architecture based on muscle movements

3.2.2 基于伪造线索的面部关键点网络构建方案

目前, 最常见的人脸伪造方案主要包括面部替换(FaceSwap)、唇形同步(Lip-sync)和人脸操纵(Puppet-master)。面部替换将视频中的人脸替换为其他人的面部特征, 通常需要对整个面部进行对齐和替换操作。唇形同步使视频中人物的唇型与预设音频同步变化, 主要伪造目标是唇部区域。人脸操纵通过建立视频中人物的 3D 模型, 使其做出指定的面部表情, 包括头部运动, 一般需要对唇部区域进行特殊处理。

从伪造痕迹的角度来看, 不同的人脸伪造方案存在不同的漏洞。通过面部替换生成的视频, 内脸和外脸在俯仰角度上存在不一致。内脸指的是视频中人物中心五官的面部特征, 而外脸则指的是人物的外轮廓。例如, 当人物低头时, 内脸和外脸的俯仰角度可能不一致, 导致伪造痕迹暴露。而对于唇形同步类的深度伪造技术, 由于人类的嘴唇运动与语音是高度相关的, 因此在进行唇形同步伪造时, 往往难以完美模拟出真实的嘴唇运动, 这也为检测提供了一定的线索。检测方法可以通过检测嘴唇张开程度和唇颌距离等指标来进行识别伪造视频。人脸操纵类的伪造容易出现头部运动和面部表情的不协调, 产生伪造痕迹; 也存在 3D 模型无法完美捕捉人物的面部特征, 导致存在伪造痕迹。

为了更加直观地观察真伪视频在各个面部连接间的差异, 图 4、图 5 和图 6 分别对唇颌距离(LM_8-LM_{57}), 唇鼻距离(LM_{33}-LM_{51})和嘴唇开闭距离(LM_{62}-LM_{66})进行了可视化。每组曲线的第一行为真实视频, 横轴表示整段视频的各个视频帧, 纵轴表示关键点间的距离。使用的视频均来自 FakeAVCeleb 数据集[21]的 FVRA 子集。通过对真伪视频中的关键面部关键点间的距离进行细致分析, 可以观察到一系列显著的异常现象。这些异常揭示了伪造视频与真实视频之间微妙的差别。

图 4 展示了唇颌距离的曲线图。真视频(用黑色线条表示)的唇颌距离曲线呈现出明显的峰谷变化, 这反映了在真实情况下人在说话或产生面部表情变化时唇部和下颌的自然运动。然而, 在假视频(以彩色线条表示)中, 这种峰谷变化变得很不明显, 甚至完全消失。这种差异表明, 伪造视频在模拟唇部和下颌的运动时可能存在技术上的不足, 无法完美地模拟真实场景的运动状态。或者伪造者为了掩盖伪造痕迹而故意平滑了这些变化。类似的现象也出现在鼻唇距离的曲线中, 如图 5 所示。在真视频中, 鼻唇距离的变化能够反映出人在不同面部表情和说话时鼻部和唇部之间的相对运动。而在假视频中, 这种变化往往被削弱或消除, 使得伪造视频的面部动作显得更为僵硬和不自然。更进一步地, 观察嘴唇开闭的曲线。在图 6 中, 真视频的嘴唇开闭曲线展现出了明显的张开和闭合动作, 这与人说话时嘴唇的自然运动相吻合。然而, 在伪造视频中, 嘴唇的开闭程度变得较为不明显, 缺乏真实视频中的那种明显的张开和闭合动作。这种差异可能是由于伪造视频在模拟嘴唇运动时的技术限制, 或者是伪造者为了避免被检测而故意降低了嘴唇运动的幅度。

针对上述分析观察, 本节在基于人脸面部肌肉运动的面部网络基础上, 添加了与伪造痕迹高度相关的关键点及其连接, 包括唇颌距离(LM_i, i = 8-57), 嘴唇开闭(LM_i, i = 62-66), 鼻唇距离(LM_i, i = 33-51),

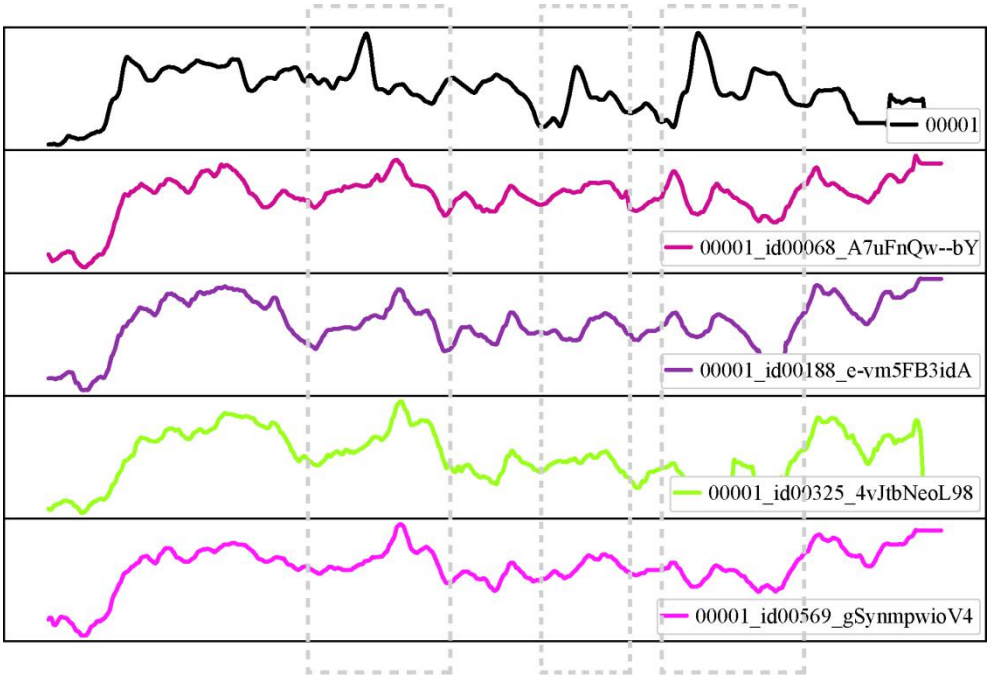


图 4　唇颌距离(LM_8-LM_{57})

Figure 4　Lip-to-Chin Distance (LM_8-LM_{57})

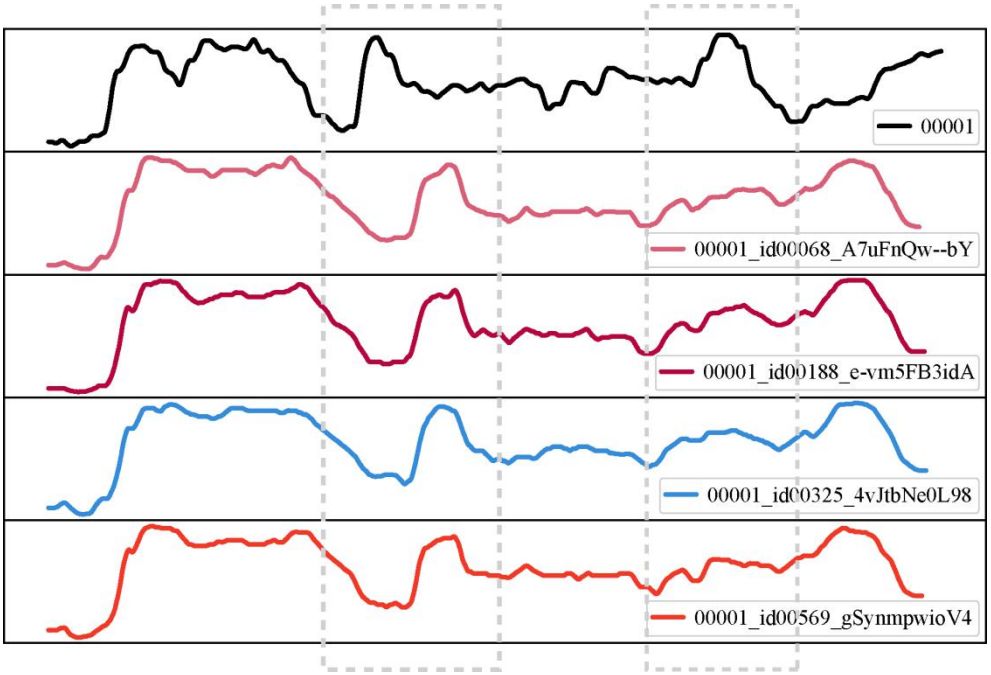


图 5　唇鼻距离 (LM_{33}-LM_{51})

Figure 5　Lip-to-Nose Distance (LM_{33}-LM_{51})

面部轮廓(LM_i, i = 0-6)。这样的面部网络构建方案充分考虑了面部替换、唇形同步和人脸操纵等常见伪造手段带来的固有伪造痕迹, 帮助网络更好地捕捉伪造线索, 实现伪造人脸的识别工作。最终基于肌肉运动和伪造线索的面部关键点网络结构设计如图 7 所示, 图中蓝色连接为基于肌肉运动的面部网络, 橙色连接为基于伪造线索的面部网络。

3.3　基于时空图注意力机制的面部关键点特征表示网络

本节在网络实现上, 引入了图注意力网络(Graph Attention Network, GAT)和门控循环单元(Gated Recurrent Unit, GRU)来实现对短时和长时伪造痕迹

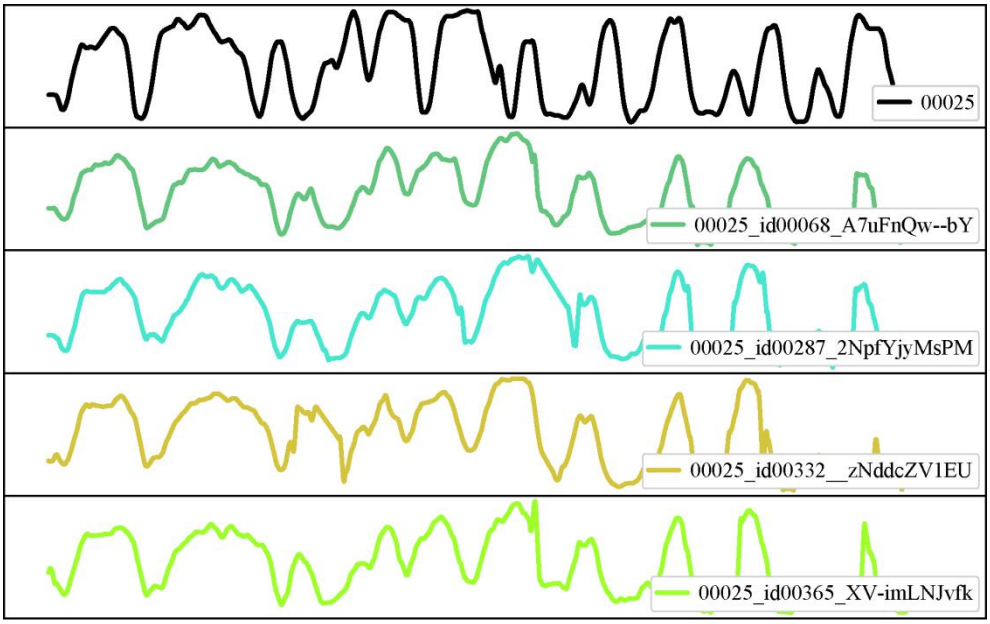


图 6　嘴唇开闭距离 (LM_{62}-LM_{66})

Figure 6　Lip Closure (LM_{62}-LM_{66})

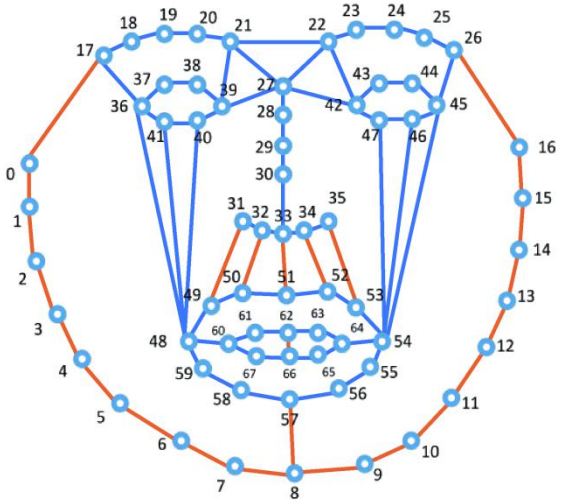


图 7　基于肌肉运动和伪造线索的面部关键点网络结构设计

Figure 7　Design of facial landmark network architecture based on muscle movements and forgery clues

的捕捉。图注意力网络使用上述构建好的面部关键点连接图作为输入，其中图的节点表示各个面部关键点，边表示节点之间的关系或相互作用(边的连接包括帧内面部关键点连接和帧间短时连接)。图注意力网络通过计算节点之间的时空关系，动态地为每个节点分配权重。这一过程通常包括三个步骤：首先，网络提取每个节点的特征；其次，通过时空信息构建节点之间的关系图；最后，应用注意力机制，根据关系图的上下文信息，学习节点间的相对重要性，以便更有效地聚合特征。这种动态学习使得网络能够灵活应对不同场景下的特征变化。

GRU 网络能够在面部关键点时空图的基础上更好地理解不同帧之间的长时连续性和时间相关性。具体来说，GRU 通过更新门和重置门来有效捕捉长期时间信息，从而更好地建模时间序列数据的动态变化。与图注意力网络的协作中，GRU 负责处理图注意网络输入的所有序列特征，而图注意力网络则利用图结构捕捉序列内部的关系。二者结合可以实现对长时和短时伪造痕迹的建模。

OpenFace 工具能够从人脸图像中提取出 68 个面部关键点。在任意给定的 t 时刻，每个面部关键点都可以通过坐标对$l_i^t = [x_i^t, y_i^t]$来表示，其中 i 的取值范围是从 1 到 68，对应着人脸上不同位置的关键点。坐标对中的x_i^t和y_i^t分别代表所有关键点在横轴和纵轴上的位置，即可以方便整合横纵坐标序列为$x_i^t = (x_1^t, x_2^t, ..., x_{68}^t)$和$y_i^t = (y_1^t, y_2^t, ..., y_{68}^t)$。从形式上看，图结构是由节点集合 V 和边集合 E 组成的，记作 $G = (V, E)$。节点集合 V 包含了一组 N 个图节点，表示为 $V = (v_1, v_2, ..., v_N)$。而边集合 E 则代表了一系列面部关键点之间的连接关系。为了更清晰地描述这些连接关系，本文引入了一个大小为 $N \times N$ 的邻接矩阵 A，其中元素$A_{i,j}$反映了节点 i 和节点 j 之间的连接强度；若两节点间无连接，则$A_{i,j}$的值为 0。在这里，为了实现对短时伪造痕迹的捕捉，使用τ表示局部序列的时

间长度。此时的时空邻接矩阵表示为$\tilde{A}_{\tau}^{t}$, 通过枚举局部时空邻域中所有可能的邻居, 此时的邻接矩阵的构成如式(3.3.1)所示。

$$\tilde{A}_{\tau}^{t}=(\tilde{A}_{(1)}^{t},\tilde{A}_{(2)}^{t},\ldots,\tilde{A}_{(\tau)}^{t})\in\Re^{\tau\times N\times N} \tag{3.3.1}$$

其中, $\tilde{A}_{\tau}^{t}$表示大小为 $\tau\times N\times N$ 的跨时间图, 通过在局部时空邻域的τ个相邻帧上建立当前时间戳的节点与其邻居之间的关系。如果以当前时间戳τ的一个节点为例, 则该节点会连接到$\tau\times N$ 个邻居。在短时面部关键点序列的处理中, 将每个视频帧的每个面部关键点的特征集合作为输入数据, 构成特征张量$X^{C\times T\times N}$。其中, C 代表每个节点的特征维度, T 代表帧数, 而 N 则代表节点数量。考虑输入序列为每个时间步的大小τ的滑动窗口, 在每个时间步生成一个局部面部关键点序列, 可表示为式(3.3.2):

$$X_{\tau}^{t}=\{x_{t-\frac{\tau}{2}:t+\frac{\tau}{2}}\in C\times\tau\times N\,|\,0\le t<T\} \tag{3.3.2}$$

为了更好地区分相似的面部运动, 通过引入注意力机制, 能够根据节点之间的时空关系动态地学习每个节点对于其他节点的重要性。首先, 对于图注意力网络的第 l 层, 其输出的特征可以表示为$h_l=(h_1^l,h_2^l,\ldots,h_N^l)$, 其中$h_i^l\in\Re^F$, N 代表图网络节点的数量, 在本文中 $N=68$。为了计算式(3.3.1)中连通边的注意力权重, 首先将节点特征转换为高维特征, 通过自注意力机制对相邻关键点的注意力分数进行计算, 计算过程如式(3.3.3)所示:

$$e_{ij}=\alpha(\phi h_i^l,\phi h_j^l) \tag{3.3.3}$$

其中, $\phi\in\Re^{F'*F}$, 将节点特征转换为高维特征; $\alpha(\cdot,\cdot)$ 为注意力机制计算函数。随后通过 softmax 和 LeakyReLU 激活, 计算最终归一化后的注意力分数 α_{ij}, 即式(3.3.4):

$$\alpha_{ij}=\mathrm{softmax}(\mathrm{LeakyReLU}(e_{ij})) \tag{3.3.4}$$

通过上述计算的注意力分数, 可以得到模型在第 $l+1$ 层的输出, 如式(3.3.5)所示。其中, $\mathbb{N}(i)$表示在$\tilde{A}_{(\tau)}^{t}$中, 与节点 i 相邻的节点。ϕ^l表示第 l 层节点特征转换权重矩阵。

$$h_i^{l+1}=\sigma\{\textstyle\sum_{j\in N(i)}\alpha_{ij}\,\phi^l h_j^l\} \tag{3.3.5}$$

多头注意力机制允许模型在多个表示子空间内学习注意力系数, 从而增强了自注意力学习过程的鲁棒性。本工作使用多头注意力机制, 具体来说, 通过采用 K 个独立的注意机制在 K 个头中执行变换, 模型能够更全面地捕捉输入数据中的不同特征, 提高了信息处理的多样性和准确性。多头注意力机制下的h_i^{l+1}层输出表示为式(3.3.6), 通过对 K 个注意力头输入的平均, 作为节点的最终输出值。如图 8 所示。

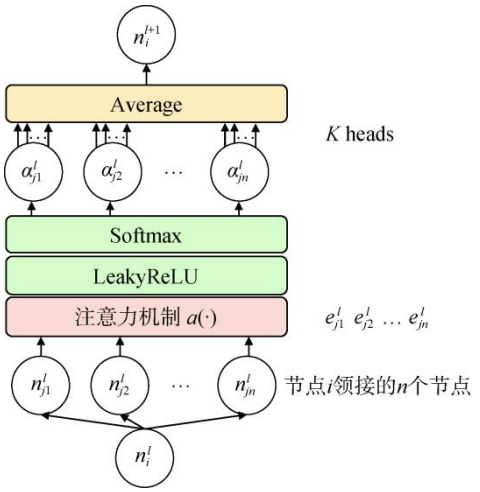


图 8　时空图注意力块框图

Figure 8　Block diagram of spatiotemporal graph attention

$$h_i^{l+1}=\sigma\{\tfrac{1}{K}\textstyle\sum_{k=1}^{K}\sum_{j\in N(i)}\alpha_{ij}^{k}\phi^l h_j^l\} \tag{3.3.6}$$

在多头注意力机制中, K 个独立的注意力机制通过将输入特征矩阵分别映射到多个子空间来执行变换。每个头使用独立的线性变换生成查询(Q)、键(Key)和值(Value), 然后计算注意力权重并加权求和。这种并行处理使得模型能够捕捉不同的特征表示和信息, 从而提高学习的表达能力。

为了实现对长时伪造痕迹的建模, 本文使用了两层双向 GRU 块来建模长期的时间信息。通过上述短时伪造痕迹的构建, 时空图注意力网络的输出特征向量可以表示为$(y_1,y_2,\ldots,y_t)$。短时特征会由两层双向 GRU 处理, 用于长时建模。输出特征 $z_i=(z_1,z_2,\ldots,z_t)\in\Re^{Q\times t}$, 其中 $Q=2\times H_0$, H_0 表示每层中隐藏单位的数量, 如式(3.3.7)所示:

$$(z_1,z_2,\ldots,z_t)=\mathrm{SeqGRU}(y_1,y_2,\ldots,y_t) \tag{3.3.7}$$

最后, 一个 softmax 层将视频面部关键点的特征表示映射到标签上进行分类。该分类器将$Z_i\in\Re^{Q*t}$向量序列映射到为真实或伪造的概率p^i。随后, 模型的训练目标为最小化其与视频标签y^i的交叉熵损失, 模型的损失函数如公式(3.3.8)所示:

$$L=-\frac{1}{N}\sum_{i=1}^{N}y^i\log p^i+(1-y^i)\log(1-p^i) \tag{3.1.8}$$

4　实验结果与分析

4.1　数据集介绍

为了对本方案在伪造检测方面的性能进行细致评估, 实验环节选用了 FaceForensics++ 数据集[3](包括 RAW, C23, C40 子集)、FakeAVCeleb 数据集[21]、FaceForensics++ social 数据集[22]。

FaceForensics++ 数据集发布于2019年, 是一个用于深度学习研究的人脸合成视频数据集。该数据集采集了1000段原始视频, 然后运用多种伪造技术生成伪造视频。FaceForensics++数据集包含了不同的合成方法和技术, 以及不同的人脸合成风格。该数据集包括1000条原始视频, 每条视频使用多种伪造手段, 包括Face2Face[23]、FaceSwap[24]、DeepFakes[25]和NeuralTextures[26], 旨在弥补传统数据集伪造质量低和伪造技术单一的缺陷。在新版数据集中, 作者添加了FaceShifter[27]伪造手段, 进一步丰富了数据集的伪造种类。

FakeAVCeleb数据集在音视频层面均进行了伪造处理。FakeAVCeleb数据集2021年发布, 不仅囊括了深度伪造的视频内容, 还包含了唇音同步的伪造音频, 为伪造检测研究提供了丰富的素材。在构建过程中, 该数据集以VoxCeleb2[28]中的500个真实视频为基础, 通过FaceSwap和FSGAN[29]等方法, 生成了多达19,500个伪造说话人脸视频。同时, 为了实现音频与视频内容的同步伪造, 还采用了SV2TTS[30]和Wav2Lip[31]等技术, 进行语音伪造生成和唇形同步处理。从音视频的伪造状态来看, FakeAVCeleb 数据集根据音视频伪造类型被划分为四个类别: 真实视频真实音频(REAL)、真实视频伪造音频(RVFA)、伪造视频真实音频(FVRA)以及伪造视频伪造音频(FVFA)。

FaceForensics++ Social 数据集[26]是通过将FaceForensics++数据集[4]的部分视频上传至Facebook和YouTube, 然后重新下载这些视频而获得的真实社交网络压缩视频数据集。该数据集仅使用了FaceForensics++ 数据集的验证集和测试集部分。FaceForensics++ Social数据集可以用于评估伪造检测方案在处理通过流行的在线社交平台分享的视频时的鲁棒性和泛化能力。

4.2 对比方法

为了全面评估本模型的性能表现, 在实验中与一系列先进的伪造检测模型进行了细致的性能比较。重点的对比对象包括LRNet[13]、FAMM[18]、SPSL[11]以及FT-two-stream[32]等模型。为确保实验结果的公正性和准确性, 上述参与比较的模型均仅从视觉角度出发, 进行特征提取和伪造检测, 未涉及任何音频信息。

1) LRNet[13]: 提出了一个高效而鲁棒的框架, 通过精确几何特征和时间建模来检测深度伪造视频。LRNet模型设计了一种新的校准模块来提高面部关键点几何特征提取的精度, 构建了双流递归神经网络(RNN)来充分利用时间特征。对于LRNet网络不同的RNN组合, 模型分为g1, g2, g1+g2三个变体。

2) FAMM[18]: 提出了一种基于面部肌肉运动框架来解决压缩深度伪造人类视频检测问题。首先从连续帧中定位人脸, 并从人脸图像中提取地标。然后, 通过对五个感官和面部区域进行建模, 利用连续的面部标志来构建面部肌肉运动特征。

3) SPSL[11]: 该方法将空域图像和相位谱结合, 捕获人脸伪造的上采样伪影, 以提高人脸伪造检测的可迁移性。并从理论上分析了利用相位谱的有效性。

4) FT-two-stream[32]: 本文通过分析压缩深度伪造视频的帧级和时间级序列, 提出了一种双流方法。利用帧级和时间级数据流提取图像特征和时间不一致特征来检测深度伪造说话人脸视频。使用FFmpeg提取关键帧作为帧级的输入。对于时间级流, 直接使用视频作为输入。

4.3 评价指标和实验安排

实验将准确率(ACC)和ROC曲线下面积(AUC)作为评估指标, 这些指标通常用于深度伪造说话人脸检测领域。准确率(ACC)反映了模型对真实和伪造样本的准确分类程度。最好的ACC是接近1, 表示分类器准确地识别了绝大多数样本。ROC曲线下面积(AUC)值是一个模型性能的综合评估指标, 越接近1, 表示模型在所有可能的阈值下都能更好地区分正样本和负样本。为后续实验安排会从算法的综合性能、多种场景下的鲁棒性、模型设计和模型可解释性多个方面进行实验。

4.4 算法综合性能评估

在表1和表2中, 展示了本方案FALNet在FakeAVCeleb数据集和FaceForensics++和上的检测性能。表格中每种伪造类型下的最优结果用黑色粗体表示。

在FakeAVCeleb数据集上, 本方案表现出色, 如表1所示。针对包含视觉伪造的FVRA(真实音频, 伪造视频)和FVFR(伪造音频, 伪造视频)子集, 本方案在各种伪造手段上均取得了98.5%以上的AUC, 表明模型能够有效地区分真伪视频。同时, 该模型在部分子集上的性能甚至超越了基于音视频的研究 AD DFD[33]。本方案在WL(FVRA)伪造手段上的AUC值显著高于其他对比方法, 达到了99.6%, 表明本方案在检测唇形伪造时具有明显优势。在FVFA伪造方法情况下, 虽然某些对比方法在某些子类别上的检测表现较好, 如AD DFD在WL(FVFA)上的AUC值为100.0%, RealForensics[14]在GAN-WL上的AUC值为99.8%, 但本方案在FVFA的整体性能上表现更为稳定。

表 1 FakeAVCeleb 数据集实验结果(AUC ↑,%)

Table 1 Experimental results on FakeAVCeleb (AUC ↑,%)

检测方案	伪造手段			
	FVRA		FVFA	
	WL	FS-WL	GAN-WL	WL
Xception	88.3	93.5	68.5	91.0
LipForensics	97.7	99.9	68.1	98.7
AD DFD	97.4	99.7	55.4	**100.0**
FTCN	97.4	**100.0**	78.3	96.4
RealForensics	93.0	99.1	**99.8**	96.7
FALNet(ours)	**99.6**	**100.0**	**99.8**	98.9

注：加粗为最优值。

表 2 FaceForensics++数据集(Raw 子集)实验结果(AUC ↑,%)

Table 2 Experimental results on the FaceForensics++ (Raw subset) (AUC ↑,%)

检测方案	伪造手段			
	DF	F2F	FS	NT
LRNet (g1)	95.2	91.1	96.5	84.0
LRNet (g2)	96.0	92.3	97.9	86.9
LRNet (g1+g2)	98.9	97.9	99.3	96.8
FAMM	95.8	95.7	97.0	91.5
SPSL	98.5	94.6	98.1	80.5
FT-two-stream	98.0	94.0	94.0	90.0
FALNet(ours)	**99.5**	**98.6**	**100.0**	**97.9**

注：加粗为最优值。

通过表 2, 可以看到本方案在 FaceForensics++数据集的各个伪造类别上均取得了优异的结果。首先, FALNet 在 FaceForensics++ 数据集的 DeepFakes、Face2Face、FaceSwap 和 NeuralTextures 四种伪造手段上的 AUC 值分别达到了 99.5%、98.6%、100.0%和 97.9%。这意味着本方案在检测这些伪造手段时表现出了非常高的准确性, 特别是在 DeepFakes 和 FaceSwap 上的性能尤为出色, 达到了接近完美的检测效果。接下来, 从表格中可以看出, 与 LRNet 方案的三种变体(g1、g2 和 g1+g2)、FAMM、SPSL 以及 FT-two-stream 相比, 本方案在四种伪造手段上的 AUC 值均显著较高, 在各个伪造类别上相比于第二名的工作, AUC 分别提高了 0.6%, 0.7%, 0.7%, 1.1%, 展现了面对多种伪造手段时良好的检测效果。

4.5 算法鲁棒性评估

为了测试本方案 H.264 重压缩下模型检测效果和鲁棒性, 本部分在 FaceForensics++数据集的压缩子集上进行实验。在表 3 展示了本方案 FALNet 在 FaceForensics++数据集不同压缩系数下(C23, C40)的检测准确率。在各个伪造类别下最优效果由黑色粗体表示。

表 3 FaceForensics++数据集(C23,C40 子集)实验结果(AUC ↑,%)

Table 3 Experimental results on the FaceForensics++ (C23,C40 subset) (AUC ↑,%)

压缩系数	检测方案	伪造手段			
		DF	F2F	FS	NT
C23	LRNet (g1)	93.4	85.5	95.0	94.7
	LRNet (g2)	94.2	90.0	96.6	96.6
	LRNet (g1+g2)	**98.9**	**98.2**	98.9	**98.9**
	FALNet(ours)	**98.9**	**98.2**	**99.3**	97.5
C40	LRNet (g1)	80.7	74.9	78.1	79.3
	LRNet (g2)	82.5	77.6	83.0	81.3
	LRNet (g1+g2)	94.6	87.1	96.8	94.3
	FAMM	90.0	91.0	92.8	85.5
	SPSL	93.5	86.0	92.3	76.8
	FT-two-stream	94.3	86.7	85.3	80.5
	FALNet(ours)	**95.7**	**93.5**	**98.2**	**97.9**

注：加粗为最优值。

首先, 在压缩系数为 C23 时, 本方案在 DeepFakes、Face2Face 和 FaceSwap 上的准确率分别达到了 98.9%、98.2%和 99.3%, 取得了最优结果, 超越了 LRNet 方案的多种变体。在 NeuralTextures 上的准确率略低于其他伪造类别, 但检测结果依然达到了 97.5%。在压缩系数为 C40(高压缩率)时, 本方案在 DeepFakes、Face2Face、FaceSwap 和 NeuralTextures 四种伪造手段上的准确率(ACC)分别达到了 95.7%、93.5%、98.2% 和 97.9%, 与其他对比方法相比, 本方案在四种伪造手段上的性能相比于次优工作, 分别提升了 1.1%、2.5%、1.4%、3.6%。这显示了本方案在高压缩率下依然能够保持出色的伪造人脸视频的检测性能。

为了进一步探究面部关键点在面临压缩时的鲁棒性, 图 9 进一步比较了在不同压缩条件下, 不同方案的性能变化情况。每个检测方案都在原始视频(RAW)上进行训练, 并在三个压缩版本的视频上进行了测试。

首先, 本方案在原始数据(RAW)上的 AUC 分数达到了 99.1%, 这一结果略低于基于像素级输入特征的 Xception[3]工作。在 C23 压缩级别下, AUC 分数为 97.5%, 相比原始数据下降了 1.6%。在 C40 压缩级别下, AUC 分数为 94.1%, 相比原始数据下降了 3.4%。尽管随着压缩级别的增加, AUC 分数有所降低, 但整

体上仍然保持了较高的性能。这显示了本方案在压缩数据上仍然具有一定的鲁棒性，能够有效地检测出伪造痕迹。与其他方法相比，尽管 Xception 方法在原始数据上的 AUC 分数略高于本方案，但在 C23 和 C40 压缩级别下，其性能下降幅度较大，分别下降了 6.4%和 13.2%。而 X-Ray[32]方法在压缩数据上的性能下降更为显著，AUC 分数在 C40 压缩级别下仅为 61.6%，下降了 37.5%。这一实验结果也表明了像素级特征在面对压缩时的不鲁棒。

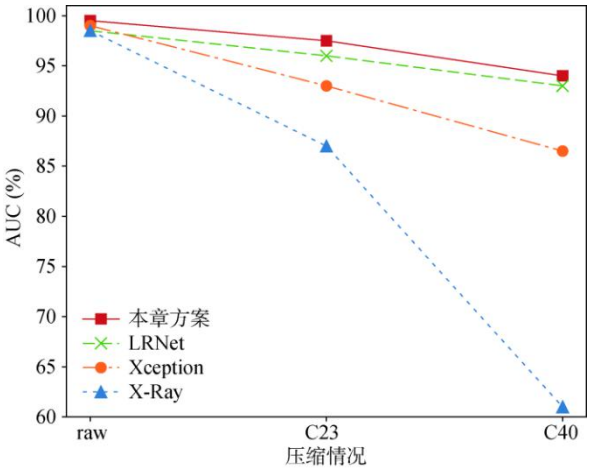


图 9　压缩情况下不同方法的 AUC 分数

Figure 9　AUC score of different methods under compression

本文方案 FALNet 和 LRNet 均采用面部关键点作为检测信号，而 Xception 和 X-Ray 则将图像空间作为检测信号。观察到，在原始数据集上，Xception 表现出最佳性能；然而，当数据集经历压缩后，Xception 的性能显著下降，基于图像空间的 X-Ray 也呈现出相似的趋势。与之相比，基于面部关键点的检测方案在压缩情况下表现更加稳定，具有更强的鲁棒性。本文方案在面对 C23 和 C40 压缩时，性能仅分别下降了 1.6%和 3.4%，远远优于基于图像空间的其他方法。这一结果突显了面部关键点作为检测信号的优势，尤其是在受压缩影响时，能够确保检测的稳定性和鲁棒性。

4.6　主流社交应用环境中的算法鲁棒性评估

为了对模型在主流社交应用平台下的检测效果和鲁棒性进行评估，本部分在 FaceForensics++ Social 数据集上进一步实验，检验本章方案的鲁棒性和泛化能力。表 4 展示了模型在 FaceForensics++ Social 数据集各伪造手段上本方案 FALNet 与其他工作性能的比较。在各个社交平台下，各个伪造类别的最优结果用黑色粗体表示，次优用下划线表示。在这一组实验中，模型在 FaceForensics++ 数据集的 C23 子集上进行训练，得到的模型在 FaceForensics++ Social 数据集的测试集上进行测试，以检验在实际场景下的分类能力，以及模型在面对不同种类压缩时的泛化能力。

表 4　FaceForensics++ Social 数据集实验结果(ACC ↑, AUC ↑,%)

Table 4　Experimental results on the FaceForensics++ Social dataset (ACC ↑, AUC ↑, %)

社交平台	检测方案	伪造手段							
		DeepFakes		Face2Face		FaceSwap		NeuralTextures	
		ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Facebook	FAMM	91.00	97.35	93.73	**96.93**	94.50	**98.72**	89.00	94.97
	FT-two-stream	92.50	97.55	82.50	89.69	90.25	95.98	73.00	81.90
	Re-network	74.44	83.81	76.75	84.63	77.75	84.86	63.75	69.91
	Capsule	92.20	**98.18**	88.02	95.49	93.30	98.58	72.28	80.99
	Mesonet	86.22	93.83	84.50	91.75	82.50	89.61	70.01	83.25
	FALNet (ours)	**97.83**	97.83	**96.74**	96.74	**97.46**	97.46	**96.39**	**96.40**
YouTube	FAMM	90.75	96.17	94.25	**97.97**	95.00	**98.82**	88.25	92.44
	FT-two-stream	93.48	**97.82**	80.50	86.80	79.25	91.76	77.00	80.85
	Re-network	74.44	84.01	73.25	79.39	68.17	73.00	67.75	74.58
	Capsule	92.74	97.71	87.22	94.71	82.78	97.53	70.70	82.11
	Mesonet	85.25	93.67	79.95	88.66	77.00	84.32	66.17	81.32
	FALNet (ours)	**97.48**	97.48	**97.84**	97.84	**97.85**	97.86	**95.70**	**95.71**

注：加粗为最优值，下划线为次优值。

可以看到本方案在大部分伪造手段的实验中，和当前其他最优方案相比，在 ACC 和 AUC 指标上均能取得前二的结果。观察到在 Facebook 平台上的性能表现。本方案在 DeepFakes、Face2Face、FaceSwap 和 NeuralTextures 四种伪造手段上的准确率(ACC)分别达到了 97.83%、96.74%、97.46%和 96.39%，同时 AUC 值也较高，这显示了本方案在 Facebook 平台上对各种伪造手段的检测能力较强。通过分析在

YouTube 平台上的性能表现可以发现，本方案在 DeepFakes、Face2Face、FaceSwap 和 NeuralTextures 上的准确率分别为 97.48%、97.84%、97.85%和 95.70%，同样保持了较高的性能。与 Facebook 平台相比，本方案在 YouTube 平台上的性能略有波动，但整体仍表现出色。

结果表明本文方案在面对通过社交平台 Facebook 和 YouTube 传输的伪造视频时，保持了良好的检测性能。进一步检验了本文方案在面临实际数据时，能保持良好的鲁棒性，同时也表明该方案不仅在受控环境下表现优异，在真实世界的复杂环境中同样具有很好的泛化能力和鲁棒性。

4.7 面部关键点网络性能评估消融实验

为了对基于肌肉运动和伪造线索的面部关键点网络性能进行评估，如表 5 所示，本部分对面部关键点网络设计方案进行了消融实验，主要测试了面部关键点网络的连接设计以及图神经网络中时序连接的引入对最终检测结果的影响。表中各种压缩类型下，对各个伪造手段的最优结果用黑色粗体显示。

实验结果表明，将肌肉运动和伪造痕迹同时考虑，并采用时序连接的方案，在大多数情况下能够取得最佳结果。仅考虑肌肉运动和时序连接时，在部分伪造手段的部分压缩类型上也能取得和充分考虑各方面因素时一样的准确率，但是整体来看，基于伪造痕迹的面部连接的加入仍然对性能有正向的帮助，使得模型在 Face2Face 的 C23 和 C40 类别上取得了 98.20%和 93.53%的准确率，相比于仅考虑肌肉运动的情况均有接近 1%的提升。这也表明，精心设计的网络结合了肌肉运动和伪造线索的特征，通过有效的信息聚合和特征提取，提高了对深度伪造的识别能力。相比之下，全连接网络并没有保留面部关键点图的拓扑结构，无法充分利用肌肉运动和伪造线索的信息，从而在深度伪造检测任务中表现不及本章设计的说话面部关键点网络，在整体性能上表现相对较差。同时，全连接的网络由于在注意力计算时，聚合了全部关键点间的信息，计算量相比面部关键点网络设计的网络更大。

表 5 面部关键点网络设计方案性能评估(ACC ↑, %)

Table 5 Performance evaluation of the facial landmark network design scheme (ACC ↑, %)

图网络结构	压缩类型	伪造手段			
		DeepFakes	Face2Face	FaceSwap	NeuralTextures
关键点全连接+时序	RAW	98.20	**98.84**	99.28	**97.85**
	C23	98.20	97.84	98.57	98.92
	C40	95.32	93.17	97.85	97.49
肌肉运动+时序	RAW	**99.64**	97.85	99.28	97.13
	C23	98.20	97.48	**99.28**	**99.28**
	C40	**95.68**	92.45	**98.21**	**98.21**
肌肉运动+伪造痕迹	RAW	**99.64**	98.56	**100.00**	96.06
	C23	98.56	97.12	**99.28**	**99.28**
	C40	94.60	93.53	97.49	96.06
肌肉运动+伪造痕迹+时序	RAW	**99.64**	98.56	**100.00**	**97.85**
	C23	**98.92**	**98.20**	**99.28**	97.49
	C40	**95.68**	**93.53**	**98.21**	97.85

注：加粗表示最优值。

同时，短时时序连接的引入对模型效果也有所提升。这一结果表明，对于深度伪造检测任务来说，捕捉短时伪造痕迹至关重要。深度伪造技术常常涉及对视频的局部区域进行修改，这些修改可能仅在视频的短暂时间内出现，例如在一帧或几帧之内。因此，引入时序连接可以帮助模型更好地捕捉这些短暂的伪造痕迹，从而提高对深度伪造的检测性能。通过结合多种线索，并引入时序连接，网络能够更好地捕捉深度伪造视频中的特征，从而提高检测的准确性和鲁棒性。

4.8 模型可解释性分析

本实验通过对图注意力网络的注意力权重进行可视化，更深入地分析模型对于伪造视频的判别过程，并探究各个关键点间的连接对最终判别结果的贡献。本实验针对 FakeAVCeleb 数据集中使用 FaceSwap 和 Wav2lip 两种不同方法生成的伪造视频

进行了可视化分析。具体结果如图 10 和图 11 所示。图中第一行为进行换脸或唇形同步的原始真实视频，第二行为经过换脸或唇形同步后的生成视频，第三行为对伪造视频关键点间的注意力权重较大的连接。

在图 10 中可以观察到，针对 FaceSwap 生成的视频 (00217_id01105_qxMlGJlyfA8_faceswap.mp4)，注意力权重集中在唇颌距离等区域。这是因为 FaceSwap 涉及将目标人脸的五官与当前人脸进行对齐，往往导致内脸与外脸间存在不一致。模型通过关注这些区域的细微差异，能够准确地捕捉到换脸操作引起的伪造痕迹。因此，在今后的模型改进中，我们可以进一步强化这些区域的特征提取策略，甚至加入多尺度特征提取方法，以增强模型对局部和全局伪造特征的捕捉能力。

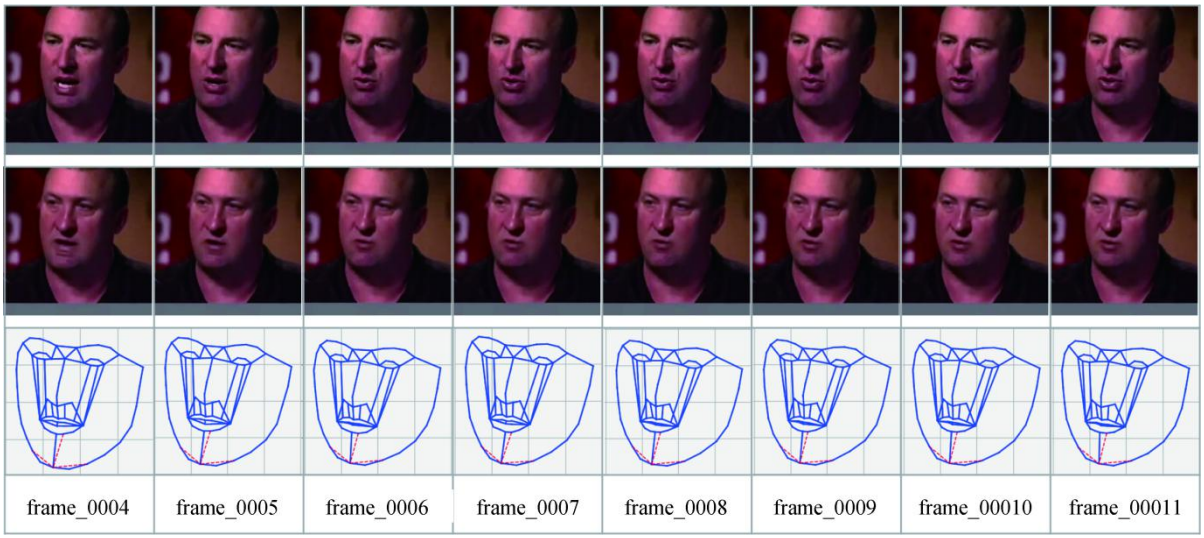


图 10　FakeAVCeleb 数据集中 FaceSwap 伪造视频注意力可视化

Figure 10　Attention visualization of FaceSwap forged videos in the FakeAVCeleb dataset

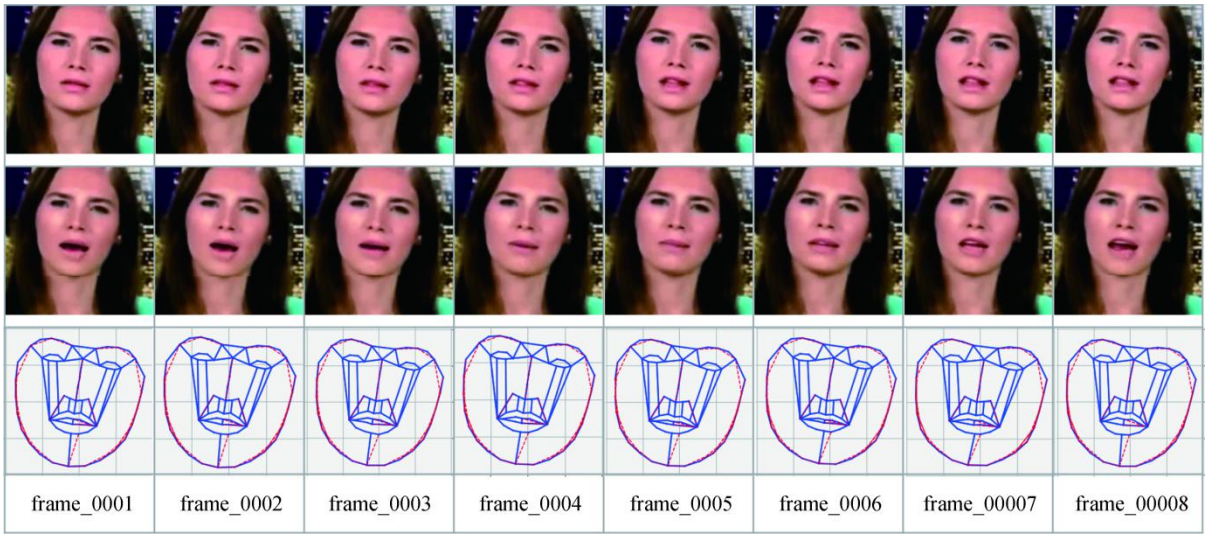


图 11　FakeAVCeleb 数据集中 Wav2lip 伪造视频注意力可视化

Figure 11　Attention visualization of Wav2lip forged videos in the FakeAVCeleb dataset

在图 11 中可以观察到，针对 Wav2lip 生成的视频(00171_id00261_wavtolip.mp4)，模型的注意力权重集中在嘴唇及其与鼻子之间的区域。这是因为音频驱动生成的唇形同步技术难以完美模拟真实唇部运动，模型通过在这些区域集中注意力，能够识别出伪造视频中的不自然运动。这一分析表明，模型在检测唇形同步伪造时的敏感区域非常明确。在未来的改进中，可以通过多头注意力机制加强模型在这些区域的响应能力，或结合音频特征，以提升模型在多模态伪造视频中的表现。

通过对模型注意力权重的可视化分析，可以观察到在本文方案的模型在决策过程中，其关注的重

点区域与实际的伪造痕迹存在的重点区域具有高度的一致性。这表明本文方案能够有效地捕捉并识别出伪造视频中的关键痕迹。此外, 这种一致性也使得本文方案的模型具备了良好的可解释性, 使得模型的决策过程更加透明且易于理解。

5 总结和展望

本文面向伪造说话人脸检测技术存在的鲁棒性问题, 基于说话行为所带来的肌肉运动以及深度伪造说话人脸视频生成过程带来的伪造线索, 提出了基于说话行为相关面部关键点的鲁棒伪造人脸检测方案。实验结果表明, 本方案在多个公开数据集的视频伪造子集上均取得了超过 98%的检测准确率, 相比于现有的先进方法, 本方案的 AUC 值取得了 0.6%~1.1%的提升, 且在压缩场景下具有良好的鲁棒性, 检测 AUC 值保持在 94%以上, 有效地提升压缩情况下伪造说话人脸检测的鲁棒性, 具有良好的应用价值。未来的研究工作可在此基础上融入音频信号, 探索音视频间的关系, 进一步提升模型在多种伪造类别上的检测性能。

参考文献

[1] Lewis A, Vu P, Duch R M, et al. Deepfake Detection with and without Content Warnings[J]. *Royal Society Open Science*, 2023, 10(11): 231214.

[2] Afchar D, Nozick V, Yamagishi J, et al. MesoNet: A Compact Facial Video Forgery Detection Network[C]. *2018 IEEE International Workshop on Information Forensics and Security*, 2018: 1-7.

[3] Rossler A, Cozzolino D, Verdoliva L, et al. FaceForensics++: Learning to Detect Manipulated Facial Images[C]. *2019 IEEE/CVF International Conference on Computer Vision*, 2019: 1-11.

[4] Bonettini N, Cannas E D, Mandelli S, et al. Video Face Manipulation Detection through Ensemble of CNNS[C]. *2020 25th International Conference on Pattern Recognition*, 2021: 5012-5019.

[5] Coccomini D A, Messina N, Gennaro C, et al. Combining EfficientNet andVision Transformers forVideo Deepfake Detection[C]. *Image Analysis and Processing – ICIAP 2022*, 2022: 219-229.

[6] Frank J, Eisenhofer T, Schönherr L, et al. Leveraging Frequency Analysis for Deep Fake Image Recognition[C]. *The 37th International Conference on Machine Learning*, 2020: 3247-3258.

[7] Qian Y Y, Yin G J, Sheng L, et al. Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues[C]. *Computer Vision – ECCV 2020*, 2020: 86-103.

[8] Miao C T, Tan Z C, Chu Q, et al. Hierarchical Frequency-Assisted Interactive Networks for Face Manipulation Detection[J]. *IEEE Transactions on Information Forensics and Security*, 2022, 17: 3008-3021.

[9] Li J M, Xie H T, Li J H, et al. Frequency-Aware Discriminative Feature Learning Supervised by Single-Center Loss for Face Forgery Detection[C]. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 6454-6463.

[10] Liu H G, Li X D, Zhou W B, et al. Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain[C]. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 772-781.

[11] Li Y Z, Chang M C, Lyu S W. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking[C]. *2018 IEEE International Workshop on Information Forensics and Security*, 2018: 1-7.

[12] Haliassos A, Vougioukas K, Petridis S, et al. Lips Don't Lie: A Generalisable and Robust Approach to Face Forgery Detection[C]. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 5037-5047.

[13] Sun Z K, Han Y J, Hua Z Y, et al. Improving the Efficiency and Robustness of Deepfakes Detection through Precise Geometric Features[C]. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 3608-3617.

[14] Masi I, Killekar A, Mascarenhas R M, et al. Two-Branch Recurrent Network for Isolating Deepfakes in Videos[M]. Computer Vision – ECCV 2020 International Publishing, 2020: 667-684.

[15] Zheng Y L, Bao J M, Chen D, et al. Exploring Temporal Coherence for More General Video Face Forgery Detection[C]. *2021 IEEE/CVF International Conference on Computer Vision*, 2021: 15024-15034.

[16] Yan Z Y, Sun P, Lang Y B, et al. Multimodal Graph Learning for Deepfake Detection[EB/OL]. 2022: arXiv: 2209.05419. https://arxiv.org/abs/2209.05419.

[17] Liao X, Wang Y M, Wang T Y, et al. FAMM: Facial Muscle Motions for Detecting Compressed Deepfake Videos over Social Networks[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, 33(12): 7236-7251.

[18] Li M, Liu B B, Hu Y J, et al. Deepfake Detection Using Robust Spatial and Temporal Features from Facial Landmarks[C]. *2021 IEEE International Workshop on Biometrics and Forensics*, 2021: 1-6.

[19] Baltrušaitis T, Robinson P, Morency L P. OpenFace: An Open Source Facial Behavior Analysis Toolkit[C]. *2016 IEEE Winter Conference on Applications of Computer Vision*, 2016: 1-10.

[20] Hu J, Liao X, Wang W, et al. Detecting Compressed Deepfake Videos in Social Networks Using Frame-Temporality Two-Stream Convolutional Network[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, 32(3): 1089-1102.

[21] Khalid H, Tariq S, Kim M, et al. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset[EB/OL]. 2021: arXiv: 2108.05080. https://arxiv.org/abs/2108.05080.

[22] Marcon F, Pasquini C, Boato G. Detection of Manipulated Face Videos over Social Networks: A Large-Scale Study[J]. *Journal of Imaging*, 2021, 7(10): 193.

[23] Thies J, Zollhöfer M, Stamminger M, et al. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos[C]. *2016 IEEE*

Conference on Computer Vision and Pattern Recognition, 2016: 2387-2395.

[24] FaceSwap. https://github.com/MarekKowalski/FaceSwap. (2016-10-14).

[25] Deepfake. https://github.com/deepfakes/faceswap. (2018-02-6).

[26] Thies J, Zollhöfer M, Nießner M. Deferred Neural Rendering: Image Synthesis Using Neural Textures[J]. *ACM Transactions on Graphics*, 2019, 38(4): 1-12.

[27] Li L Z, Bao J M, Yang H, et al. Advancing High Fidelity Identity Swapping for Forgery Detection[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 5073-5082.

[28] Chung J S, Nagrani A, Zisserman A. VoxCeleb2: Deep Speaker Recognition[EB/OL]. 2018: arXiv: 1806.05622. https://arxiv.org/abs/1806.05622.

[29] Nirkin Y, Keller Y, Hassner T. FSGAN: Subject Agnostic Face Swapping and Reenactment[C]. *2019 IEEE/CVF International Conference on Computer Vision*, 2019: 7183-7192.

[30] Jia Y, Zhang Y, Weiss R J, et al. Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis[C]. *The 32nd International Conference on Neural Information Processing Systems*, 2018: 4485-4495.

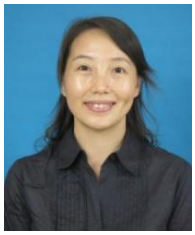
[31] Prajwal K R, Mukhopadhyay R, Namboodiri V P, et al. A Lip Sync Expert Is all You Need for Speech to Lip Generation in the Wild[C]. *The 28th ACM International Conference on Multimedia*, 2020: 484-492.

[32] Li L Z, Bao J M, Zhang T, et al. Face X-Ray for More General Face Forgery Detection[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 5000-5009.

[33] Sung C S, Chen J C, Chen C S. Hearing and Seeing Abnormality: Self-Supervised Audio-Visual Mutual Learning for Deepfake Detection[C]. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023: 1-5.

黄逸焕 CCF 学生会员, 于 2022 年在武汉大学网络空间安全专业获得工学学士学位。现在武汉大学国家网络安全学院网络空间安全专业攻读博士学位。研究领域为多媒体内容安全。研究兴趣包括: 音视频伪造检测。Email: yihuanhuang@whu.edu.cn

彭荔 于 2024 年在武汉大学网络空间安全专业获得工学硕士学位。研究领域为多媒体内容安全。研究兴趣包括: 音视频伪造检测。Email: pengli29@whu.edu.cn



任延珍 CCF 会员, 于 2009 年在武汉大学通信与信息系统专业获得博士学位。现任武汉大学国家网络安全学院教授。研究领域为多媒体内容安全。研究兴趣包括: AI 交互安全, 多媒体取证, 多媒体伪造检测, 多媒体信息隐藏及隐写分析等。Email: renyz@whu.edu.cn

王丽娜 CCF 会员, 于 2001 年在东北大学获得博士学位。现任武汉大学国家网络安全学院教授, 博士生导师。研究领域为系统安全、多媒体信息隐藏、隐写分析理论和技术、网络安全、云安全及人工智能安全。Email: lnwang@whu.edu.cn