中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2025)07-2343-21

论文引用格式:Yao W D, Li P C, Zhao Y and Wu H C. 2025. Review of research on face deepfake detection methods. Journal of Image and Graphics, 30(7):2343-2363(姚文达, 李盼池, 赵娅, 吴洪超. 2025. 人脸深度伪造检测方法研究综述. 中国图象图形学报, 30(7):2343-2363)[DOI:10.11834/jig.240586]

人脸深度伪造检测方法研究综述

姚文达[1],李盼池[1],赵娅[1*],吴洪超[2]

1. 东北石油大学计算机与信息技术学院(网络空间安全学院),大庆　163318; 2. 大庆市公安局网络警察分局,大庆　163318

摘　要:深度伪造技术是一种基于深度学习的合成技术,旨在生成高度逼真的合成图像、音频或视频,包括人脸伪造、声音模仿和人体姿态合成等内容。其中,人脸深度伪造可以实现非常逼真的换脸效果,广泛应用于电影、动画制作等领域。然而,该技术的滥用导致不雅视频与虚假新闻的传播,带来了恶劣的社会影响。为应对这些负面影响,众多学者提出一系列检测方法,以有效识别伪造人脸图像或视频。然而,当前的检测方法类型多样、优缺点各异,且应用场景各不相同,鉴于此,本文对相关研究进行了系统的归纳与整理。首先,整理了人脸深度伪造检测常用数据集与评价指标;其次,从图像级伪造人脸检测与视频级伪造人脸检测两个领域出发,根据特征选择的不同,将前者划分为基于空间域和基于频率域的检测方法,将后者划分为基于时空不一致、基于生物特征和基于多模态的检测方法,并详细总结了各类方法的原理、优缺点及发展趋势,特别地,考虑到文本生成图像/视频的流行以及生成式人工智能在多模态创作上的显著进步,综述了针对文本生成图像/视频的检测方法和基于多模态的检测方法;最后,梳理了人脸深度伪造检测领域的研究现状及瓶颈,并对未来的研究与发展方向进行了探讨。

关键词:深度伪造检测;人脸伪造检测;人脸图像;人脸视频;空间域特征;频率域特征;时序特征;多模态特征

Review of research on face deepfake detection methods

Yao Wenda[1], Li Panchi[1], Zhao Ya[1*], Wu Hongchao[2]

1. School of Computer and Information Technology(School of Cyber Science and Technology), Northeast Petroleum University, Daqing　163318, China; 2. Cyber Police Sub-Bureau, Daqing Public Security Bureau, Daqing　163318, China

Abstract: Deepfake technology refers to the synthesis of images, audio, and videos by using deep learning algorithms. This technology enables the precise mapping of facial features or other physical characteristics from one person onto a subject in another video, achieving highly realistic face-swapping effects. With advancements in algorithms and the increased accessibility of computational resources, the threshold for utilizing deepfake technology has gradually lowered, bringing convenience and numerous social and legal challenges. For example, deepfake technology is used to bring deceased actors back to the screen, providing a novel experience to the audience. Meanwhile, it is frequently exploited to impersonate citizens or leaders for fraudulent activities, produce pornographic content, or create fake news to influence public opinion. Consequently, the importance of deepfake detection technology is increasing, making it a significant focus of current

收稿日期:2024-10-26;**修回日期**:2024-12-20;**预印本日期**:2024-12-27

*通信作者:赵娅　zhaoya@nepu.edu.cn

基金项目:国家自然科学基金项目(62471124);黑龙江省自然科学基金项目(LH2022F006);黑龙江省教育科学规划重点项目(GJB1421114);教育部产学合作协同育人项目(220501867161323);东北石油大学青年引导性基金项目(2022YDL-14)

Supported by: National Natural Science Foundation of China (62471124); Natural Science Foundation of Heilongjiang Province, China (LH2022F006); Key Projects of Heilongjiang Provincial Education Science Planning (GJB1421114); Ministry of Education Industry-University Cooperation Collaborative Education Program (220501867161323); Northeast Petroleum University Youth Guidance Fund (2022YDL-14)

research. To detect images and videos synthesized via deepfake technology, researchers must design models that can uncover subtle traces of manipulation within these media. However, accurately identifying these traces remains challenging due to several factors that complicate the detection process. First, rapid advancements in deepfake technology have made differentiating fake images and videos from authentic content increasingly difficult. As techniques such as generative adversarial networks (GANs) and diffusion models continue to evolve and improve, the texture, lighting, and motion within synthesized media become more seamlessly realistic, imposing significant challenges on detection models that seek to recognize subtle cues of manipulation. Second, forgers can employ a variety of countermeasures to obscure traces of manipulation, such as applying compression, cropping, or noise addition. Furthermore, forgers may create adversarial samples that are specifically crafted to exploit and bypass the vulnerability of detection models, making the identification of deepfake even more complex. Thirdly, the generalizability of deepfake detection methods remains a significant hurdle, because different generative techniques leave behind distinct forensic traces. For example, GAN-generated images frequently exhibit prominent grid-like artifacts in the frequency domain, while images produced through diffusion models typically leave only subtle, less detectable traces in this domain. Therefore, detection models that do not exclusively rely on low-level, technique-specific features but instead focus on capturing deep, generalized features that ensure robustness and applicability across diverse forgery types and detection scenarios are crucial. To address these multifaceted challenges, numerous scholars have proposed a variety of detection methods that are designed to capture nuanced traces left by deepfake manipulations. For example, certain approaches focus on identifying subtle forgery artifacts within the frequency domain of images, capitalizing on the distinct spectral anomalies that forgeries frequently introduce. Other methods prioritize assessing temporal consistency across video frames, because unnatural transitions or frame-level inconsistencies can indicate synthesized content. In addition, some detection strategies focus on evaluating synchronization among different modalities within videos, such as audio and visual elements, to detect inconsistencies that may reveal forgery. At present, several review papers in academia have summarized key research and developments within this domain. However, given the rapid advancements in generative artificial intelligence (AI), fake faces created with diffusion models have recently gained popularity, with scarcely any review that addresses the detection of such forgeries. Furthermore, as generative AI progresses continues advancing toward multimodal integration, deepfake detection methods are similarly evolving to incorporate features from multiple modalities. Nonetheless, the majority of existing reviews lack sufficient focus on multimodal detection approaches, underscoring a gap in the literature that this review seeks to address. To provide an up-to-date overview of face deepfake detection, this review first organizes commonly used datasets and evaluation metrics in the field. Then, it divides detection methods into image-level and video-level face deepfake detection. Based on feature selection approaches, image-level methods are categorized into spatial-domain and frequency-domain methods, while video-level methods are categorized into approaches based on spatiotemporal inconsistencies, biological features, and multimodal features. Each category is thoroughly analyzed with regard to its principles, strengths, weaknesses, and developmental trends. Finally, current research status and challenges in face deepfake detection are summarized, and future research directions are discussed. Compared with other related reviews, the novelty of this review lies in its summary of detection methods that specifically targets text-to-image/video generation and multimodal detection methods. This review is aligned with the latest trends in generative AI, offering a comprehensive and up-to-date summary of recent advancements in face deepfake detection. By examining the latest methodologies, including those developed to address forgeries created through advanced techniques, such as diffusion models and multimodal integration, this review reflects the ongoing evolution of detection technology. It highlights the progress made and the challenges that remain, positioning itself as a valuable resource for researchers who aim to navigate and contribute to the cutting-edge developments in this rapidly advancing field. A comprehensive analysis of face deepfake detection methods reveals that current techniques achieve nearly 100% accuracy within the training datasets, particularly those leveraging advanced models, such as Transformers. However, their performance frequently declines significantly in cross-dataset testing, particularly for spatial-domain and frequency-domain detection methods. This decline suggests that these approaches may fail to capture essential, generalizable features that are robust across varying datasets. By contrast, biological feature-based methods demonstrate superior generalization capabilities, successfully adapting to different contexts. However, they require carefully tailored training data and specific application conditions to reach optimal per-

formance. Meanwhile, multimodal detection methods, which integrate features across multiple modalities, offer enhanced robustness and adaptability due to their layered approach. However, this added complexity frequently results in higher computational costs and increased model intricacy. Given the diversity in feature selection, along with the unique advantages and limitations inherent to each detection approach, no single method has yet provided a fully comprehensive solution to the deepfake detection challenge. This reality underscores the critical need for continued research in this evolving field and highlights the importance of this review in mapping current advancements and identifying future research directions.

Key words: deepfake detection; face forgery detection; face image; face video; spatial domain feature; frequency domain feature; temporal features; multimodal features

0 引 言

生成式人工智能(artificial intelligence generated content, AIGC)因在图像生成、自然语言处理等方面取得显著进展而备受关注(刘安安 等,2024)。许多先进的生成模型,如 GPT(generative pre-trained Transformer)、StyleGAN(style-based generative adversarial network)等,能通过自监督学习等方法从大量未标记的数据中训练,捕捉数据中的潜在模式并进行创作。目前,研究者正探索将不同类型的数据结合,进行多模态创作,这种能力促进了视频生成、虚拟试衣等应用的发展。

随着 AIGC 技术的发展,深度伪造(deepfake)技术也不断进步。深度伪造是利用深度学习,尤其是生成对抗网络(generative adversarial network, GAN)生成逼真图像或视频的技术,它通过替换原始影像中的人脸、表情或声音实现以假乱真。当前的深度伪造方法主要基于 GAN、自动编码器和扩散模型,GAN 通过生成器和判别器的对抗训练,生成逼真的伪造内容;自动编码器通过数据重构生成伪造内容;扩散模型通过向一幅真实图像逐步添加噪声,使其转化成纯噪声图像,再从纯噪声图像开始逐步去噪,逐渐得到目标图像,在此过程中,模型在输入文本或图像的引导下,学习到从噪声生成图像的能力。

目前,深度伪造技术可分为面部重演(Song 等,2022a; Zhang 等,2021)、面部编辑(Xu 等,2022b; Jiang 等,2021)、人脸替换(Xu 等,2022a)、GAN 生成(Xia 等,2021; Karras 等,2021)和扩散模型生成(Chandra 等,2024; Mehta 等,2024; Ren 等,2024)。自 2017 年以来,该技术发展迅速,网络上相继出现 FaceSwap(Korshunova 等,2017)、Face2Face(Thies 等,2019b)等面向大众的一键换脸工具。然而,深度伪造技术的快速发展带来许多负面影响。因此,研究人脸深度伪造检测对确保信息真实性、减少金融诈骗和网络犯罪具有重要意义。

当前,深度伪造检测技术已取得显著进展,且随着 AIGC 在多模态生成上的进步,深度伪造检测也正朝着多模态特征融合的方向发展。本综述介绍了本领域的常用数据集与评价指标,并根据各类方法的特征选择不同,对现有方法进行了分类,更新了人脸深度伪造检测的最新代表性方法,总结了各方法的优缺点及发展趋势。最后,讨论了本领域现有瓶颈和未来研究方向。与同领域的其他综述相比,本综述总结了针对文本生成图像/视频的检测方法及相关数据集,并整理了基于多模态的检测方法。为帮助理解人脸深度伪造检测的技术演变和关键时间节点,本综述整理了本领域的里程碑事件时间线,如图 1 所示。

1 数据集与评价指标

1.1 数据集

高质量的人脸深度伪造检测数据集对提升模型性能至关重要。人脸深度伪造检测数据集中的真实数据通常从网络上采集或请演员拍摄,伪造数据由各种伪造技术生成。以下是对常用数据集的总结。

1)CelebA(https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html)。Liu 等人(2015)发布大型人脸属性数据集 CelebA,其中包含超过 20 万幅名人真实图像,每幅图像包含 40 种标签及 5 个面部关键点,广泛用于人脸识别、生成等领域。该数据集虽然不属于伪造人脸数据集,但其数据量大、标注齐全的特性有助于训练 GAN 模型生成逼真的图像,进而帮助检测模型提高识别 GAN 卷积痕迹的能力。然而,该数据集中的图像标签准确率不高,限制了其在精细化研究

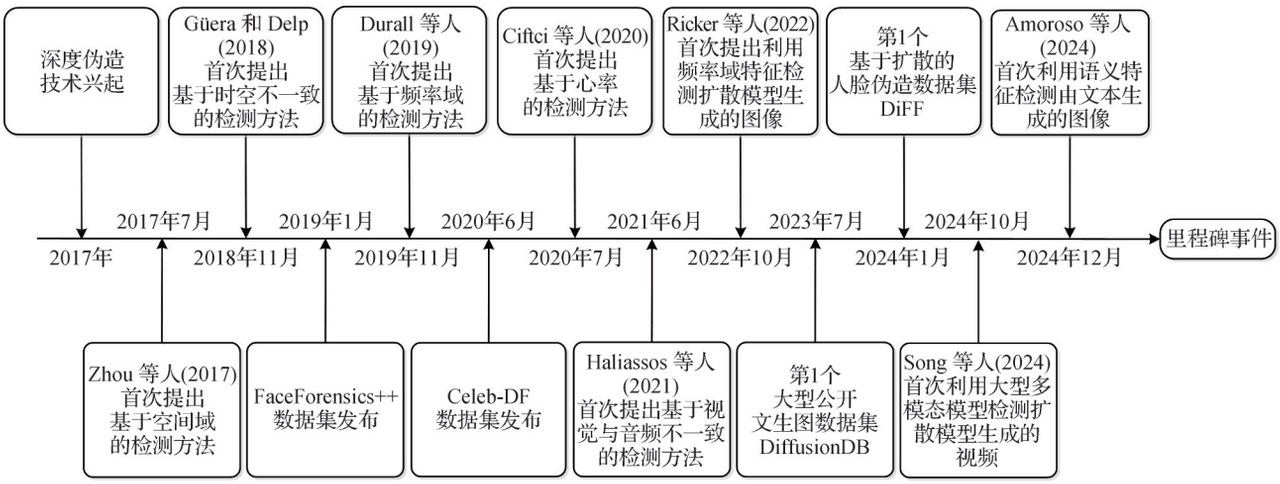

图 1　人脸深度伪造检测领域里程碑事件时间线

Fig. 1　Timeline of milestone events in face deepfake detection

中的应用。

2）DeepfakeTIMIT（https://www.idiap.ch/en/scientific-research/data/deepfaketimit）。Korshunov 和 Marcel（2018）基于 VidTIMIT 数据集创建了 DeepfakeTIMIT 数据集，其中包含 320 个高质量视频和 320 个低质量视频，可以评估检测方法对不同质量视频的适应性，并且包含音频轨道，可以用于开发基于音频的检测方法。该数据集的伪造视频由基于 GAN 的软件生成，因此伪造种类较为单一，适合基础模型的训练。

3）FaceForensics++（https://github. com/ondyari/FaceForensics）。为解决数据集质量低、伪造类型少等问题，Rössler 等人（2019）基于 FaceForensics 数据集（Rössler 等，2018）构建了规模更大、内容更丰富的 FaceForensics++数据集，该数据集包含 1 000 个真实视频，并通过 Deepfakes、Face2Face、FaceSwap 和 NeuralTextures（Thies 等，2019a）4 种技术生成了 5 000 多个伪造视频，还采用 H. 264 codec 压缩方式对所有视频进行了原始、高质量和低质量 3 种不同程度的压缩。该数据集适用于伪造人脸检测、伪造区域定位、伪造方法识别以及数字媒体取证等多种场景，因规模大、伪造技术多以及标注完整而成为最广泛使用的数据集，但存在视频合成效果明显的缺点。

4）DFDC（https://www. kaggle. com/c/deepfake-detection-challenge/data）。Dolhansky 等人（2020）构建了目前最大的公开可用的深度伪造检测数据集 DFDC（deepfake detection contest），共包括 128 154 个视频，其中真实与伪造视频的比例约为 1∶5，涵盖多种伪造技术。该数据集是在真实自然环境中录制的，涵盖走动、强光等多种复杂场景，且分辨率较高并已经过预处理。该数据集更适用于贴近真实场景的检测，其缺点是人脸占画面比例较小、在走动时人脸边界伪影比较明显。

5）Celeb-DF（https://github. com/yuezunli/celeb-deepfakeforensics/tree/master/Celeb-DF-v1）。Li 等人（2020d）构建了高质量的 Celeb-DF 数据集，其中包含 590 个真实视频和 5 639 个高质量伪造视频，涉及 59 位性别、种族和年龄分布广泛的公众人物。该数据集的伪造技术精良，避免了不自然的面部对齐、边缘伪影等瑕疵，因而常用于跨库测试中，以衡量模型对高质量伪造视频的泛化能力。但存在正负样本不均衡、数据来源单一以及规模小的问题。

6）FakeAVCeleb（https://github. com/DASH-Lab/FakeAVCeleb）。Khalid 等人（2021）创建了视听多模态数据集 FakeAVCeleb，其中包含 500 个真实视频和 19 500 个伪造视频，分为真音频和真视频、假音频和真视频、真音频和假视频、假音频和假视频 4 种不同类型。该数据集通过手动筛选确保了数据集的质量，适用于多模态深度伪造检测场景。其缺点是缺乏动态场景，音频质量有限。

7）LAV-DF（https://github.com/ControlNet/LAV-DF）。为解决现有数据集不包含内容篡改型深度伪造视频的问题，Cai 等人（2022）构建并应用了 LAV-DF（localized audio visual deepfake）数据集。与现有数据集主要关注视觉篡改不同，LAV-DF 专注于内容层面的篡改，例如将“支持”换为“反对”。该数据集包含 99 873 个假视频和 36 341 个真实视频，其中假

视频包括假音频和真视频、真音频和假视频、假音频和假视频3种伪造类型。该数据集有助于推动内容篡改型伪造检测技术发展，但存在分辨率不够高的缺点。

8）DiffusionDB（https://github.com/poloclub/diffusiondb）。DiffusionDB数据集（Wang等，2023b）是文本生成图像领域首个公开的大规模数据集，其中包括约200万幅由Stable Diffusion模型根据文本提示生成的图像，还包括180万个独特的文本提示，用以描述图像内容和风格。DiffusionDB拥有规模大、生成内容丰富、公开可用的优点，但由于文本来自不同的真实用户，该数据集存在图像质量不一致的缺点。

9）DiffusionForensics（https://github.com/Zhendong-Wang6/DIRE）。为应对现有检测方法在识别由扩散模型生成的伪造图像时面临的挑战，Wang等人（2023a）构建了名为DiffusionForensics的数据集。此数据集的真实图像来自FFHQ（flickr-face-HQ）（Karras等，2021）等公开真实图像数据集，伪造图像由Stable Diffusion等8种扩散模型生成。该数据集标注详细且生成模型多样，能广泛应用于对扩散模型生成的图像的检测中。

10）DiFF（https://github.com/xaCheng1996/DiFF）。基于扩散的数据集普遍关注一般的图像生成，缺少由扩散模型生成的伪造人脸数据集，为解决这一问题，Cheng等人（2024）发布了第1个基于扩散模型的伪造人脸数据集DiFF（diffusion facial forgery）。该数据集使用20 000多个精心收集的文本提示和10 000个视觉提示生成了超过50万幅高质量图像，涵盖13种最先进的扩散技术，且每幅图像都注明了所采用的伪造方法和相应提示。作为该领域的首个数据集，它填补了现有数据集在基于扩散的人脸伪造检测方面的空白，为该领域的研究提供了重要的数据基础。

11）总结。由于扩散模型生成图像技术的新颖性，目前用于扩散模型生成图像检测的数据集较少，相关研究通常使用扩散模型自主构建数据集，但随着扩散模型生成图像检测技术的发展，未来预计会有更多大规模、标准化的基于扩散的伪造人脸数据集公开。

与传统的计算机视觉任务不同，人脸深度伪造检测数据集处于持续更新状态，主要原因在于伪造方法不断更新，数据集需要不断补充新的伪造数据以适应这些变化。但数据集中的伪造人脸普遍是批量生成的，而伪造者伪造的人脸往往是经过多次优化的，二者质量差异较大（曹申豪 等，2022），且数据集中人工标注的标签不免具有主观性与不准确性。因此，创建高质量的数据集仍是未来研究的重点。

1.2　评价指标

评价指标是衡量检测模型性能的重要工具，它不仅能衡量模型的准确度，还能直观反映模型的鲁棒性和在各种场景下的适用性。现有研究普遍将深度伪造检测归为二分类问题，因此通常将准确率（accuracy，ACC）和ROC（receiver operating characteristic curve）曲线下的面积（area under the curve，AUC）作为评价指标。

然而，在AIGC迅速发展、深度伪造检测方法日益复杂的情况下，只使用ACC和AUC作为评价指标无法全面评价检测方法的表现，应从数据需求量、模型复杂度、检测精度和泛化能力等多方面建立一个全面的评价标准，因此，新评价标准的构建也是未来研究的重要方向。

2　图像级伪造人脸检测

图像级伪造人脸检测是指通过检测图像或视频帧以判断其内容是否真实，根据特征选择的不同，可分为基于空间域的检测方法和基于频率域的检测方法。图2所示为图像级伪造人脸检测方法的总览图。

2.1　基于空间域的检测方法

基于空间域的检测方法是指在图像的原始像素空间中进行特征提取和处理的方法。部分研究者直接分析图像像素值（Nataraj等，2019）、局部边缘（Li等，2020a；冯才博 等，2024）或纹理（Hu等，2021；朱新同 等，2021）特征，提取伪造痕迹并训练分类器。此外，也有研究者训练深度学习模型，直接将原始图像作为输入，通过深层神经网络自动学习和提取特征，从而实现有效检测。

2.1.1　基于传统图像取证的检测方法

早期的人脸深度伪造检测主要基于传统的图像取证，传统图像取证包括对图像像素的分析、对图像特征（如噪声模式、颜色分布）的统计检验等，这些技术旨在识别图像处理过程中留下的痕迹，因而可将这种方法应用于深度伪造检测。Matern等人（2019）发现深度伪造的人脸在眼睛、鼻子及牙齿区域存在

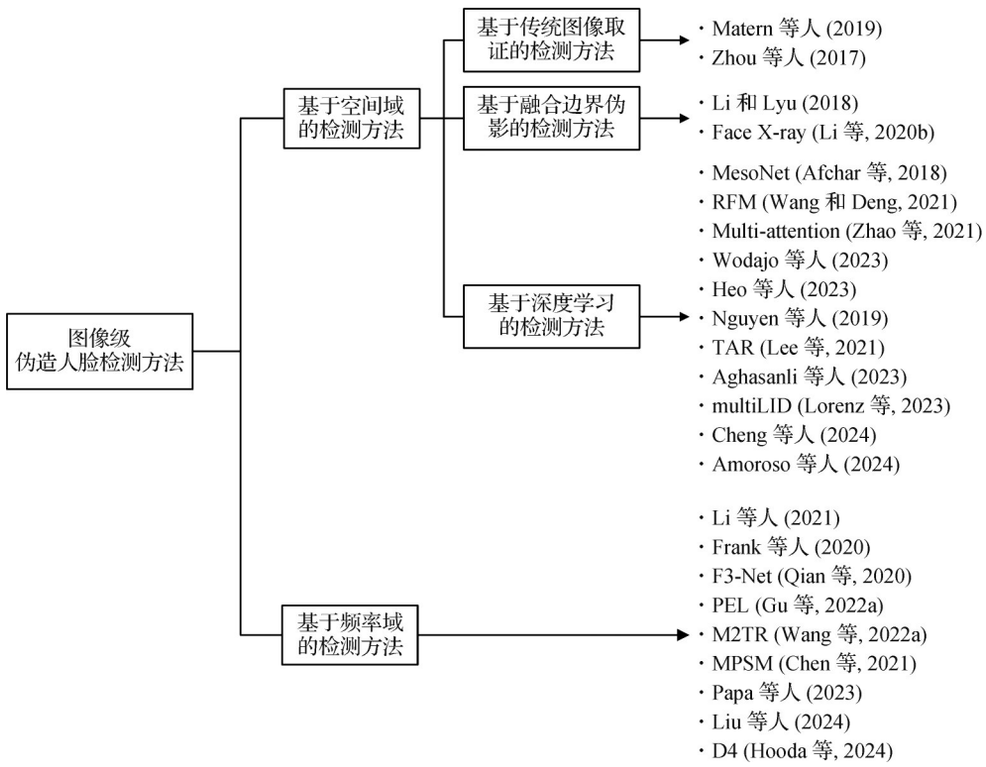

图2　图像级伪造人脸检测方法总览

Fig. 2　Overview of image-level face deepfake detection methods

颜色不一致和异常光照反射，因而构建简单的分类特征并使用逻辑回归等模型进行分类。该方法易于理解，但闭眼和牙齿不可见等情况限制了其泛化能力。Zhou等人（2017）设计了一个双流网络，其中一个流使用GoogLeNet检测面部篡改的痕迹，另一个流捕捉局部噪声残差和相机特性。最后综合两个流的检测分数做出最终判断。该方法提高了检测的准确性，但对微小的篡改不敏感且计算成本高。

2.1.2　基于融合边界伪影的检测方法

除了基于传统的图像取证，部分研究还利用脸部融合边界的伪影进行检测。Li和Lyu（2018）认为，生成的人脸在经过几何变形以匹配源视频的人脸时会留下伪影，因而采用仿射变换、高斯模糊模拟这些伪影特征，并使用ResNet-50（residual network-50）实现分类，但其准确性有待提高。Li等人（2020b）提出Face X-ray方法，该方法通过噪声分析和误差水平分析识别融合边界区域的不连续性，然后生成反映融合边界的灰度图，并将灰度图作为监督信号训练卷积神经网络（convolutional neural network，CNN）。训练后的CNN能输出反映输入图像混合边界的灰度图。其检测流程如图3所示。

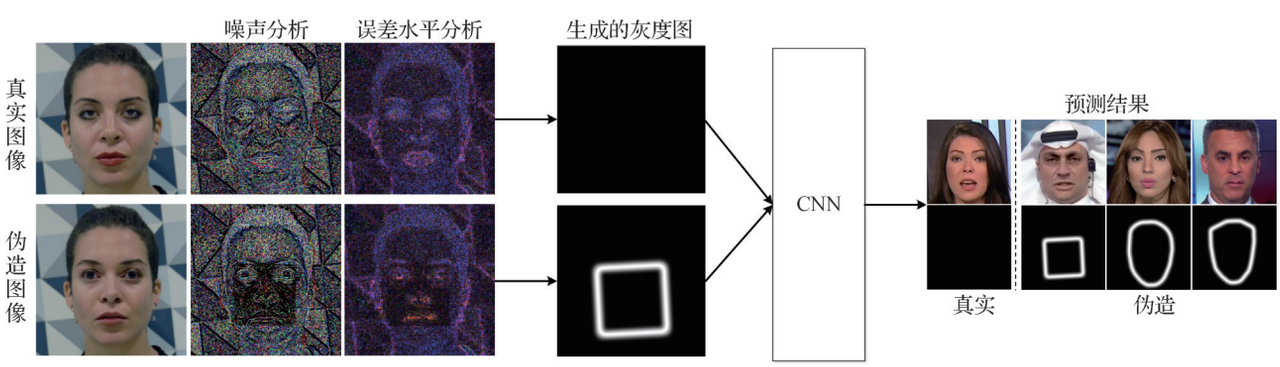

图3　Face X-ray方法检测流程图（Li等，2020b）

Fig. 3　Flow chart of Face X-ray method (Li et al., 2020b)

2.1.3　基于深度学习的检测方法

相比于手动设计算法进行特征提取，部分研究者更青睐于利用深度学习模型自动学习特征并分类。早期的方法直接将图像应用于CNN及其变种，如Afchar等人(2018)针对Deepfakes和Face2Face两种伪造技术，提出名为MesoNet的轻量级神经网络，该模型通过Inception模块优化卷积网络，以低计算成本实现了检测，但由于对伪造痕迹的挖掘不充分，该方法的泛化性较差。

在深度学习中，注意力机制能帮助模型学习深层的细微特征，因而研究者们经常应用注意力机制优化模型的泛化效果。Wang和Deng(2021)提出基于注意力的数据增强框架RFM(representative forgery mining)，通过追踪和遮挡最敏感的面部区域，引导检测器扩大注意力范围，挖掘先前被忽视的伪造特征。Zhao等人(2021)提出多注意力(muti-attention)检测网络，如图4所示。该网络通过多个空间注意力头定位伪影，并聚合浅层纹理和高级语义特征。尽管该方法在细微伪影检测中表现出色，但存在易对特定伪造方法过拟合的问题。

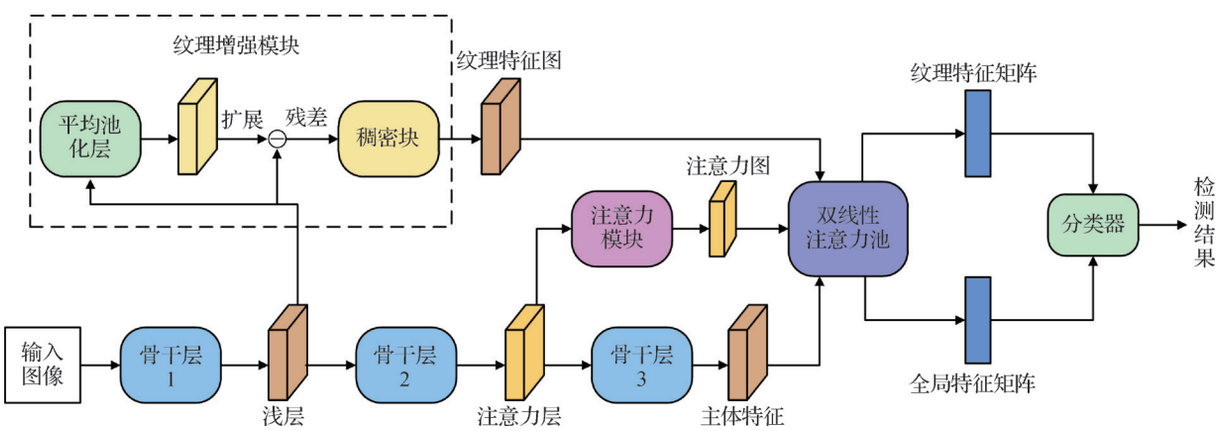

图4　多注意力检测网络结构(Zhao等,2021)

Fig. 4　Multi-attention detection network structure (Zhao et al., 2021)

随着ViT(vision Transformer)模型在计算机视觉领域取得突出进展，许多研究将ViT及其改进模型应用于深度伪造检测。Wodajo等人(2023)利用改进的ViT提取图像特征，将其与自编码器和变分自编码器生成的潜在特征结合，实现对图像内容的全面分析。Heo等人(2023)提出一种应用知识蒸馏的改进ViT模型，该模型融合patch embedding特征和CNN特征，并利用EfficientNet-B7作为教师网络进行蒸馏训练，提高了模型的泛化能力和准确性。基于ViT的方法虽然性能较为优异，但模型复杂度高且需要大量训练数据与计算资源，难以满足实际应用中的轻量级、实时性等要求，仍需针对特定任务进行优化。

针对基于深度学习的检测模型需要大量训练数据的问题，部分研究者提出解决方案。Nguyen等人(2019)将VGG-19(Visual Geometry Group 19-layer network)提取的空间特征输入到胶囊网络进行检测，由于胶囊网络的特性，即使只有少量训练数据，此方法也能展现出较好性能，但胶囊网络也存在计算量较大的问题。迁移学习是解决数据需求问题的重要手段之一，Lee等人(2021)提出一个基于迁移学习的自编码模型TAR(transfer learning-based autoencoder with residuals)，该模型通过多级迁移学习，平均检测准确率达到98.01%。多级迁移学习只需要少量新领域的数据即可实现模型更新，为降低数据需求提供了新思路。迁移学习技术虽然能在一定程度上减少模型对新领域数据集的依赖，但对源域的数据集质量要求较高，且同样存在计算复杂度高的问题。

自2022年以来，以GLIDE(guided language to image diffusion for generation and editing)为代表的扩散模型在文生图、图生图领域取得显著进展，许多研究开始关注扩散模型生成的人脸图像的检测。Aghasanli等人(2023)在ImageNet-1K数据集上对ViT模型进行预训练，并在自定义的数据集上进行微调，利用提取的特征在多个简单分类器上实现了精准分类。这表明，扩散模型生成的图像在颜色、纹理等低层次特征上存在缺陷，能被视觉模型有效

识别。

真实图像在特征空间中呈现自然的密度分布，而扩散模型生成的图像则较为分散。基于此，Lorenz等人(2023)提出multiLID(multi local intrinsic dimensionality)方法，通过估计特征空间的局部密度检测伪造图像。multiLID是由多个局部内在维度(local intrinsic dimensionality，LID)值构成的特征向量，提供细致的局部信息。作者利用ResNet提取图像的特征图表示，在提取的特征图上计算multiLID分数，最终通过随机森林进行分类。该方法思路新颖、准确率高，但泛化性有待提高。为提高检测模型的泛化能力，Cheng等人(2024)提出边缘图正则化方法。该方法利用Sobel算子提取图像的边缘信息并构建边缘图，将边缘图作为正则化项添加到目标函数中，以促进模型关注边缘特征，避免只依赖于颜色、纹理等低级特征。此外，作者还构建了基于扩散模型的人脸伪造数据集DiFF，以支持该方法在多样化数据上的测试与验证。

Amoroso等人(2024)进一步探讨了语义特征在检测中的作用，作者训练了两个线性投影层，将浅层特征与语义特征投影到不同的特征空间，并在语义空间中使用线性分类器对语义特征进行分类。实验结果表明，生成图像和真实图像在语义空间中可被有效区分，证明了语义特征可以作为检测扩散模型生成的图像的依据。Țânțaru等人(2024)认为不仅要区分图像的真假，还应输出定位图以标识被操纵的区域。作者将深度伪造定位任务视为弱监督定位问题，提出基于局部评分的定位方法。该方法采用Xception网络提取特征图，通过1×1卷积层将特征图投影到每个小块的评分。在训练阶段，结合图像级分类损失和每个小块的定位损失优化模型，最终生成检测分数和定位图。

2.1.4　基于空间域的检测方法总结

基于空间域的检测方法主要利用图像的空间信息，早期的方法主要依赖于对图像像素和浅层特征的分析，但由于缺乏高层语义信息，其泛化能力有限。基于融合边界伪影的方法具有代表性，且较为通用，但对低质量图像敏感。随着深度学习的快速发展，基于深度学习的方法成为主流，由于模型结构复杂且能自动学习特征，这些方法的检测性能大幅提高，尤其是在细微伪影的识别方面表现优异。然而，这些方法通常需要大量的数据和计算资源，且容易出现过拟合现象。

综合上述总结，本文对基于空间域的检测方法的未来研究有如下建议：对于提升模型的泛化能力，可结合迁移学习(Ranjan等，2020)和知识蒸馏(Yang等，2023a)等技术，降低对大规模数据集的依赖；或应用数据增强技术(Mandelli等，2022；Sun等，2021b，2022；Chen等，2023；李颖 等，2023)，拓展训练数据的多样性。对于提升模型的特征提取能力，可在融合边界伪影的检测中引入局部注意力机制(戴昀书 等，2023)；或采用自监督学习方法(Zhuang等，2022；Fung等，2021；Khormali和Yuan，2024)，通过设计边界预测等任务增强模型发现边界伪影的能力；或通过融合不同尺度的图像特征(Cao等，2022)提高对伪影的检测效果。最后，应推进轻量化模型设计以适应实时应用需求。表1为基于空间域的检测方法的优缺点总结。

2.2　基于频率域的检测方法

基于频率域的检测方法是指通过离散傅里叶变换(Durall等，2019)等方式将图像从空间域转换到频率域，在频率域中捕捉因篡改而引入的异常的高频信号(Li等，2021)或不规则的频率分布(Zhang等，2019)的方法。

GAN在生成高分辨率图像时，其上采样操作可能会在图像频率域中留下周期性的网格状痕迹(亦称“指纹”)。部分研究基于这种痕迹进行检测。Frank等人(2020)使用离散余弦变换(Lu等，2009)将图像转换到频率域，利用CNN模型和线性模型实现了高准确率的检测，但泛化性有待提升。

Qian等人(2020)提出F3-Net(frequency in face forgery network)框架，其结构如图5所示，该框架包括两个基于Xception的频率感知分支：频率感知分解分支用于学习频率域中的微小伪造模式；局部频率统计分支用于提取描述真实与伪造图像之间频率统计差异的高级语义信息。这两个分支通过交叉注意力模块实现特征融合。该模型对图像压缩等扰动具有较强的鲁棒性，但泛化能力较弱。

为进一步提高检测性能，研究者们将频率域特征与空间域特征结合，以更有效地捕捉图像特征并增强检测精度。Gu等人(2022a)提出渐进增强学习(progressive enhancement learning，PEL)框架，该框架通过双流网络同时处理RGB和离散余弦变换后的细粒度频率信息。每个流中均嵌入自增强模块以捕

表 1　基于空间域的检测方法优缺点总结

Table 1　Summary of the advantages and disadvantages of spatial domain-based detection methods

方法	优点	缺点	实验数据集	实验结果/%	跨数据集	实验结果/%
Matern 等人(2019)	易懂、简单易训练	泛化性差	FaceForensics	AUC=86.60	–	–
Zhou 等人(2017)	融合多种特征	对微小篡改不敏感	自建库	AUC=92.70	自建库	AUC=85.40
Li 和 Lyu(2018)	训练高效	准确性有待提高	DeepfakeTIMIT	AUC=93.20	–	–
Face X-ray(Li 等,2020b)	不依赖于伪造模型	依赖拼接痕迹	FaceForensics++	AUC=98.52	Celeb-DF	AUC=80.58
MesoNet(Afchar 等,2018)	网络层数少	只针对两种伪造	自建库	ACC=98.40	–	–
RFM(Wang 和 Deng,2021)	准确率高	对超参数设置敏感	Celeb-DF	AUC=99.97	–	–
Muti-attention(Zhao 等,2021)	能捕捉细微痕迹	可能出现过拟合	FaceForensics++	AUC=99.29	Celeb-DF	AUC=67.44
Wodajo 等人(2023)	性能优异	计算资源需求大	DFDC	AUC=99.90	–	–
Heo 等人(2023)	通过蒸馏增强泛化	参数数量巨大	DFDC	ACC=97.80	–	–
Nguyen 等人(2019)	数据量需求低	计算量大	FaceForensics++	AUC=99.37	Celeb-DF	AUC=57.50
TAR(Lee 等,2021)	少样本、高性能	训练过程复杂	FaceForensics++	ACC=99.80	自建库	ACC=89.50
Aghasanli 等人(2023)	预测结果可解释	计算成本高	自建库	ACC=99.70	自建库	ACC=84.00
multiLID(Lorenz 等,2023)	少量样本即可训练	可迁移性差	自建库	ACC=100.00	–	–
Cheng 等人(2024)	易于集成	跨域检查效果差	DiFF	ACC=99.99	–	–
Amoroso 等人(2024)	结合语义层面特征	检测过程复杂	自建库	AUC=99.68	自建库	AUC=99.68

注:“–”表示未进行跨库测试。

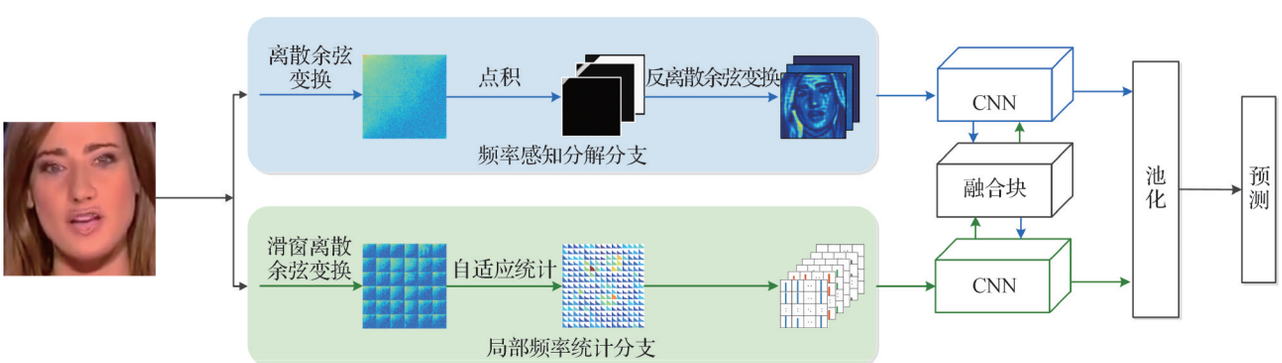

图 5　F3-Net 框架结构(Qian 等,2020)

Fig. 5　Structure of F3-Net framework(Qian et al. , 2020)

获不同输入空间中的痕迹,同时,利用互增强模块促进特征融合,实现了有效挖掘并放大细微伪造痕迹。Wang 等人(2022a)提出 M2TR(multi-modal multi-scale transformer)模型,采用双流架构,其中 RGB 流在 RGB 域的多个尺度上捕获图像不同区域之间的不一致,频率流采用可学习的频率滤波器提高抗压缩能力。该方法尽管有效,但过拟合于特定数据集。

针对模型容易对某种特定伪影过度拟合的问题,Chen 等人(2021)提出基于局部关系学习的 MPSM(multi-scale patch similarity module)框架,该框架包括一个用于衡量局部特征相似性的多尺度补丁相似性模块和一个用于融合 RGB 和频率信息的注意力模块。这些模块共同增强了模型的泛化能力。然而,该方法的计算复杂度较高且训练数据的需求量较大。

与 GAN 生成的图像不同,扩散模型生成的图像在频率域通常不呈现明显的网格状痕迹,因此,在 GAN 生成的图像上训练的检测器无法有效检测扩散模型生成的图像。然而,Ricker 等人(2022)发现扩散模型生成的图像在频率域中也表现出微小的特

定模式，且高频区域的频谱密度明显低于真实图像，可将这些特征作为分类依据。例如，Papa等人(2023)通过平均大量生成图像的频谱图得到生成模型的指纹，并利用VGG16将输入图像的频谱特征与模型指纹进行对比并分类。Liu等人(2024)观察到扩散模型生成的图像通常因过度平滑而对高斯噪声的容忍度更高，因此利用噪声残差单元得到噪声残差图像，它反映了图像对高斯噪声的响应程度，然后，将输入图像和噪声残差图像编码并融合，送入分类器进行分类。该方法简单高效，但存在过拟合的问题。

如果扩散模型生成的图像在某个频率成分上易受对抗扰动的影响，攻击者就可以通过扰动此频率成分欺骗整个检测模型。针对这一问题，Hooda等人(2024)提出D4(disjoint diffusion deepfake detection)方法，该方法将输入图像的频谱进行划分并分配到不同的检测模型中，最后通过投票机制集成检测结果。在这种情况下，攻击者需要找到多个模型的非鲁棒频率并进行扰动，增加了攻击难度。同时，多个模型并行检测提高了检测效率。

基于频率域的检测方法的优势在于能有效揭示空间域中难以察觉到的细微伪造痕迹，特别是GAN和扩散模型生成图像时可能遗留的痕迹，且对压缩、噪声等更鲁棒。但此类方法仍存在缺点，如对特定生成模型敏感、泛化能力差、计算复杂度高以及频率特征难以与其他模态信息(生理信号、音频信号)有效融合等。

综合上述总结，本文对基于频率域的检测方法的未来研究有如下建议：对于提高泛化能力，可利用自监督或无监督学习方法；或在不同数据集进行迁移学习；或应用注意力机制(Binh和Woo，2022)、特征增强(Li等，2022；Agarwal等，2021)等技术。对于降低计算复杂度，可改进模型使其更轻量化。最后，应促进频率域特征与其他模态特征的融合(Lewis等，2020)，增强模型的多样化检测能力。表2为基于频率域的检测方法的优缺点总结。

表2　基于频率域的检测方法优缺点总结

Table 2　Summary of the advantages and disadvantages of frequency domain-based detection methods

方法	优点	缺点	实验数据集	实验结果/%	跨数据集	实验结果/%
Li等人(2021)	特征学习能力强	过拟合于训练集	FaceForensics++	AUC=93.40	Celeb-DF	AUC=79.30
Frank等人(2020)	简单高效	泛化性差	CelebA	ACC=99.91	–	–
F3-Net(Qian等，2020)	能挖掘细微痕迹	依赖特定伪造类型	FaceForensics++	AUC=98.10	Celeb-DF	AUC=65.17
PEL(Gu等，2022a)	渐进式特征增强	计算复杂度高	FaceForensics++	AUC=97.63	Celeb-DF	AUC=69.18
M2TR(Wang等，2022a)	提取多尺度特征	泛化性不强	FaceForensics++	AUC=99.51	Celeb-DF	AUC=68.20
MPSM(Chen等，2021)	特征提取全面	计算复杂度高	FaceForensics++	AUC=99.46	DFDC	AUC=76.53
Papa等人(2023)	简单、易于理解	依赖高质量图像	自建库	ACC=100.00	–	–
Liu等人(2024)	方法新颖且高效	只针对扩散模型	DiffusionForensics	AUC=100.00	DiffusionDB	AUC=73.21
D4(Hooda等，2024)	鲁棒性强且高效	性能依赖于模型框架	自建库	ACC=93.00	–	–

注："–"表示未进行跨库测试。

3　视频级伪造人脸检测

视频级伪造人脸检测是指对视频中连续多帧人脸的伪造内容进行识别和检测，这种检测技术的重点不仅在于单帧图像的真实性验证，还要考虑整个视频在时序上的动作连贯性与逻辑一致性。通常，此类检测关注时序特征、生理信号、行为特征和音画同步等，通过多维特征协同分析提高检测精度。图6所示为视频级伪造人脸检测方法总览图。

3.1　基于时空不一致的检测方法

由于视频的伪造操作是逐帧进行的，因此伪造视频会出现帧与帧之间不一致的现象。基于时空不一致的检测方法利用这一特点，通过分析视频帧之间的时间序列特征以及帧内的空间特征，识别出伪造痕迹(Güera和Delp，2018)。

Sabir等人(2019)提出一种结合面部对齐和循环神经网络(recurrent neural network，RNN)的伪造

图 6　视频级伪造人脸检测方法总览

Fig. 6　Overview of video-level face deepfake detection methods

检测的方法。首先，使用 CNN 提取对齐后的面部特征，再使用一种特殊的 RNN 捕捉时序信息。这种直接将 CNN 和 RNN 结合的方法为检测时空不一致性提供了一种基本思路。类似地，Chen 等人（2020）提出一个名为 FSSpotter（face-swapped spotter）的框架，它由一个空间特征提取器（spatial feature extractor，SFE）和一个时间特征聚合器（temporal feature aggregator，TFA）组成。SFE 负责提取单帧更有辨别力的空间特征。TFA 负责捕捉连续帧之间的时间不一致性。该方法性能十分优异，但可能会对特定形状的伪造区域过拟合。

为提高泛化能力，Zheng 等人（2021）提出一种端到端框架，首先利用全时间卷积网络限制对空间伪影的学习而强调对时序不一致性的学习。随后，将学习到的特征输入到 Transformer 编码器中以捕捉长时间跨度内的不一致性。该方法的泛化能力强，但并不高效，仍需进一步优化。

Gu 等人（2022b）认为相邻帧的局部区域存在更细微的不一致，因此提出基于片段动态不一致的学习方法，将视频分割成包含连续帧的片段，利用片段内不一致模块关注相邻帧之间的局部动态不一致并通过卷积提取时序特征，然后利用片段间交互模块提取跨片段的动态信息。该方法结合片段内外不一致信息，嵌入 2D 卷积神经网络进行特征提取与分类，在多个数据集上表现出优越性能。为了更充分利用伪造视频的人脸局部时空不一致，Zhang 等人（2022）提出名为 STDT（spatiotemporal dropout Transformer）的方法，通过随机丢弃部分帧和面部补丁，在减少计算复杂度的同时充分探索了人脸的局部时空不一致性。为了进一步降低计算复杂度，Xu 等人（2024）提出 TALL-Swin（thumbnail layout-swin Transformer）方法，如图 7 所示，该方法通过密集采样提取视频片段，随后随机选择 4 个连续帧并在每帧的固定位置进行掩码以增强模型泛化能力。接着，调整这些帧为子图像并按预定义布局重组为缩略图，保留空间和时序依赖性的同时降低了计算复杂度，最后，将缩略图与 Swin Transformer 结合，利用自注意力机制提取伪造痕迹。

基于时空不一致的检测方法综合时间和空间两个维度的不一致性，适用于大多数视频场景和伪造类型。其优点在于帧间特征与图像空间特征结合，特征信息更全面。但局限性是计算复杂度高，且易受到视频后期处理影响。

综上，本文对基于时空不一致的检测方法的未来研究有如下建议：对于降低计算复杂度，可通过模型量化（卓文琦 等，2023）、剪枝等操作使模型更加轻量化，同时保持检测性能。对于提升鲁棒性，可结合增强学习和自适应策略，或针对伪造视频的对抗样本开发防御机制（Chen 等，2022；Gandhi 和 Jain，2020），进一步提升模型的鲁棒性。此外，应结合其他模态（Lewis 等，2020），利用不同信息源的互补优势，提升检测能力。表 3 为基于时空不一致的检测方法的优缺点总结。

3.2　基于生物特征的检测方法

深度伪造视频无法精准模拟人类的生理机能，例如心率、呼吸频率和面部几何运动等。基于生物特征的检测方法利用此特性，分析视频中人物的生理信号（Stefanov 等，2022）与行为特征（Li 等，2018），

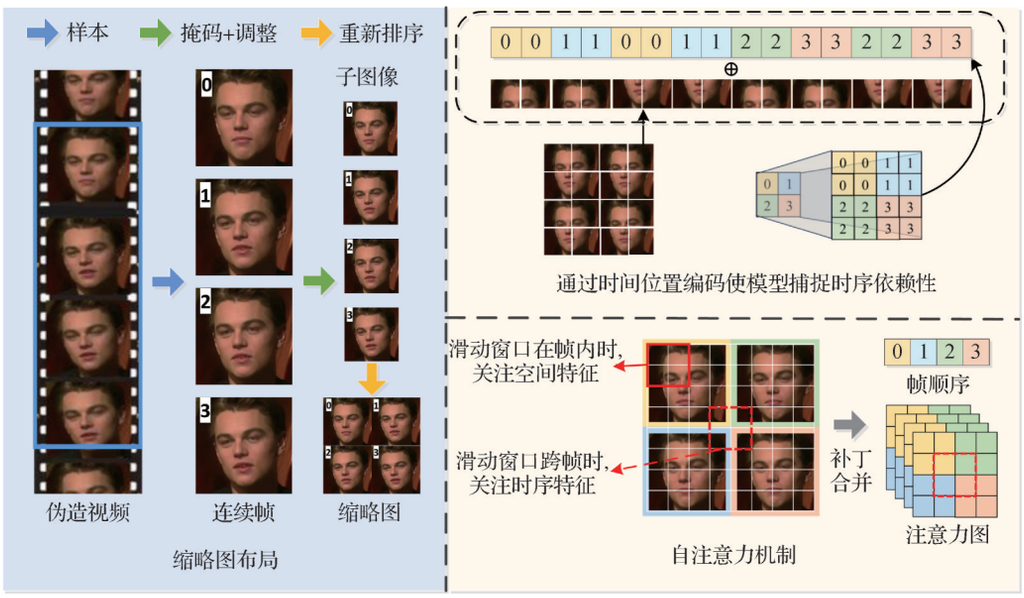

图7　TALL-Swin方法原理(Xu等,2024)

Fig. 7　Schematic of the TALL-Swin method (Xu et al., 2024)

表3　基于时空不一致的检测方法优缺点总结

Table 3　Summary of the advantages and disadvantages of spatiotemporal inconsistency-based detection methods

方法	优点	缺点	实验数据集	实验结果/%	跨数据集	实验结果/%
Sabir等人(2019)	分类性能好	计算复杂度高	FaceForensics++	AUC = 99.60	–	–
FSSpotter(Chen等,2020)	特征提取能力强	过拟合于方形伪造	FaceForensics++	AUC = 100.00	Celeb-DF	AUC = 76.26
Zheng等人(2021)	灵活、无需预训练	计算复杂度高	FaceForensics++	AUC = 99.70	Celeb-DF	AUC = 86.90
Gu等人(2022b)	特征提取能力强	难以处理长视频	FaceForensics++	AUC = 98.93	Celeb-DF	AUC = 77.65
STDT(Zhang等,2022)	时空线索捕捉高效	预处理复杂	FaceForensics++	AUC = 99.80	Celeb-DF	AUC = 69.78
TALL-Swin(Xu等,2024)	高效且泛化性强	空间信息损失	FaceForensics++	AUC = 99.87	Celeb-DF	AUC = 90.79

注:“–”表示未进行跨库测试。

识别视频中人物的真伪。

人体皮肤对光的吸收程度会因心脏跳动而产生周期性变化,这导致皮肤颜色的变化呈周期性。光电容积描记(photoplethysmographic,PPG)技术可以从皮肤捕捉这种周期性变化,并反映在PPG信号中。Ciftci等人(2020)认为深度伪造视频在生成过程中会不可避免地扰乱面部颜色的周期性变化,因此利用PPG技术提取特征并进行检测,取得了较高的准确率。这种方法成为基于生物特征的检测方法的基本思路之一。基于这一思路,Hernandez-Ortega等人(2020)提出DeepFakesON-Phys检测模型,该模型由Chen和McDuff(2018)提出的心率估计模型DeepPhys改进而来,该模型使用卷积注意力网络,包含提取帧间差异的运动模型和由注意力机制引导的外观模型,二者协同使模型关注与心率相关的特征。但该模型容易受光照等因素影响,且并未充分利用时间信息。Mao和Yang(2021)提出一种基于PPG信号和自回归(autoregressive, AR)系数的检测方案,该方案从脸部不同子区域中提取PPG信号和AR系数,将其分别建模为特征图,分别输入到结合不对称卷积结构的DenseNet中训练,最后通过特征融合提高检测准确率,但在人物运动时,其检测效果会受到影响。

Miao等人(2022)通过面部重演和口型同步合成不同的伪造视频,实验结果显示合成视频的真实感低于原始视频,这表明行为特征(如头部、嘴唇的运动)可以作为区分深度伪造视频与真实视频的有效依据。Saif等人(2024)提出一种基于面部几何结构和图卷积网络的检测方法,如图8所示,首先构建

节点表示面部特征点、边表示空间关系的图结构，并通过相同特征点的跨帧连接捕捉时间特征。最后使用图卷积网络处理图结构数据。该方法通过学习数据模式而非图像差异，显著减少了模型参数量并缩短了训练时间，并且在分类性能和泛化性方面表现出色。

伪造人脸视频本质上是对视频中的人的伪造，因此最有效的检测方法是从人体特征的角度判断视频中人物的真伪。由于各种深度伪造技术普遍难以伪造生物特征，因而此类方法通常具备更强的泛化能力。然而，此类方法也存在不足，例如，生物特征的提取通常会受到光照、遮挡和分辨率等因素的影响，从而导致检测效果不佳；此类模型较为复杂且计算复杂度高，难以满足实时检测需求；此类方法对数据集的要求较高，且生物特征的标注难度较大。

综合上述总结，本文对基于生物特征的检测方法的未来研究有如下建议：对于提升鲁棒性，可采用对抗训练的方法，生成针对性伪造样本以增强模型的鲁棒性。对于降低复杂度，可结合低复杂度网络架构，提高模型的实时性。对于数据集方面的问题，可以研究小样本学习（Aneja 和 Nießner，2020）或增量学习方法，使模型在少量标注数据下有效训练。最后，可根据个体生物特征的差异进行个性化建模，提升检测的针对性。表 4 为基于生物特征的检测方法的优缺点总结。

3.3　基于多模态的检测方法

伪造视频中往往同时包含视觉、音频和文本等多个模态的特征，而伪造视频往往会打破多模态特征之间的平衡，因此基于多模态的检测方法通过探究视频多模态特征之间的关系以识别视频真伪（Haliassos 等，2021，2022）。

某些伪造视频只篡改一小部分就足以改变整体的含义，而先前的检测方法常误判此类视频为真实视频。为解决这一问题，Cai 等人（2022）提出一种多模态检测方法，通过监督对比学习捕捉视频与音频之间的同步性，使用交叉熵损失训练分类器以识别特征，并融合图像与音频特征生成边界图，从而定位视频中的伪造段落。在真实视频中，唇部与音频之间应保持良好的同步，而在伪造视频中这种同步往往会被破坏。因此，Shahzad 等人（2022）提出一种基于唇部运

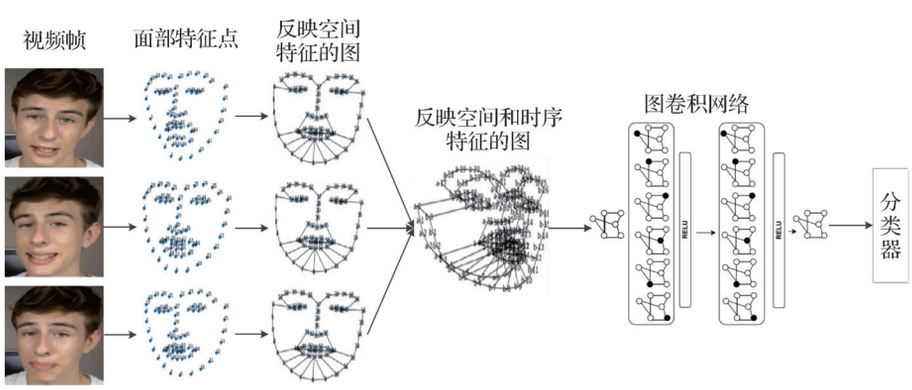

图 8　Saif 等人（2024）方法原理

Fig. 8　Schematic of the method proposed by Saif et al.（2024）

表 4　基于生物特征的检测方法优缺点总结

Table 4　Summary of the advantages and disadvantages of biological feature-based detection methods

方法	优点	缺点	实验数据集	实验结果/%	跨数据集	实验结果/%
Ciftci 等人（2020）	较为通用	对视频质量敏感	FaceForensics++	ACC = 91.07	Celeb-DF	ACC = 86.48
DeepFakesON-Phys（Hernandez-Ortega 等，2020）	通过知识迁移减少训练时间	对光照敏感	DFDC	AUC = 98.20	–	–
Mao 和 Yang（2021）	与空间特征结合	对运动敏感	FaceForensics++	ACC = 96.13	Celeb-DF	ACC = 86.57
Saif 等人（2024）	精度高、泛化性强	对遮挡与噪声敏感	FaceForensics++	AUC = 99.00	Celeb-DF	AUC = 95.00

注：“–”表示未进行跨库测试。

动的检测方法，采用了类似于Siamese网络的共享权重模型结构，通过对比从视频中提取的唇部序列与由音频合成的唇部序列之间的不一致来训练伪造检测器。然而，该方法对唇部遮挡和远距离人脸敏感。

以上方法利用了音频和视觉信号的联系，但是由于模态之间的异质性，简单的信号拼接或相减无法充分利用视听信号的之间的关系，两模态之间的关系仍应深入探索。为提高检测精度，研究者开始关注不同模态之间的信息交互与对齐，以此检测不一致性。Yang等人（2023b）提出AVoiD-DF（audio-visual joint learning for detecting deepfake）方法，如图9所示，该方法通过多模态联合解码器融合视觉与音频特征，利用双交叉注意力机制实现模态间的信息交互，确保对齐，随后采用自注意力机制揭示模态内部的伪造痕迹，并利用跨模态分类器检测音视频信号的不一致性。

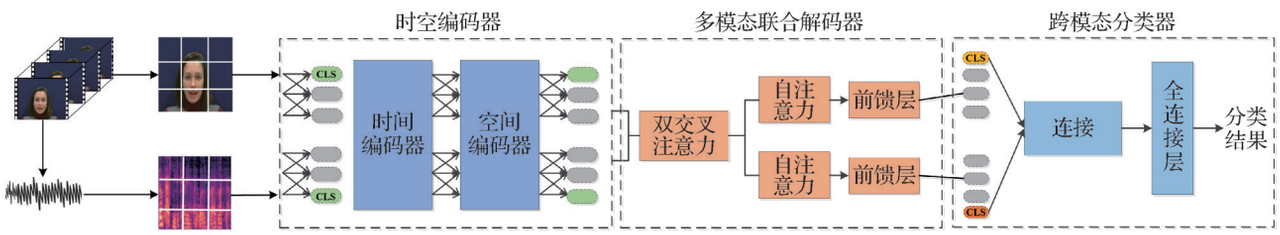

图9　AVoiD-DF方法结构（Yang等，2023b）

Fig. 9　Structure of AVoiD-DF method (Yang et al., 2023b)

为更充分地学习视觉与音频模态之间的关系，Oorloff等人（2024）提出结合自监督学习与微调的AVFF（audio-visual feature fusion）方法，在自监督学习阶段，互补编码策略促使模型更深入地学习模态之间的互补信息。这种方法减少了对标注数据的依赖，但无法检测单模态视频。为了增强模型的鲁棒性，Yin等人（2024）提出基于异构图和图神经网络的细粒度检测方法。该方法利用异构图同时捕获模态内和模态间的关系，并通过自注意力机制生成增强的特征表示，最终通过图神经网络处理得到全局表示，从而能够识别两个模态的真实性并具体分类伪造类型，但其泛化性仍有待提升。

由于生成过程依赖文本的指导，扩散模型生成的伪造视频包含文本模态的信息，因此，Song等人（2024）结合视觉特征与文本信息进行检测。首先通过大规模多模态模型将视频的视觉特征和文本特征融合为多模态特征表征，然后，通过帧内与帧间注意力机制捕捉帧内与帧间的伪造痕迹，最后，采用动态融合策略将多模态特征与时空特征融合，经过分类层得到最终的伪造预测结果。

基于多模态的检测方法综合视觉、音频和文本等特征，显著提升了伪造内容检测的准确性，并有效减少了对单一模态的依赖，具有较强的鲁棒性与泛化性。但此类方法复杂度高，预处理复杂，依赖高质量数据集，模型对模态间的复杂关系难以深入理解，且当音频和视频同时被修改时，检测效果可能降低。

综上，本文对基于多模态的检测方法的未来研究有如下建议：首先，推动自监督学习（Haliassos等，2022；Feng等，2023）的应用，减少对大量标注数据的依赖，提升模型的泛化能力；其次，进一步探索深度学习框架下多模态信息交互机制，设计更高效的特征融合方法，增强不同特征的融合效果，促进模型深入理解模态间的关系；最后，可对不同模态进行细粒度分析以提高检测精度，例如，通过分析面部细微表情变化和音频中的情感变化，识别更加隐蔽的伪造特征。表5为基于多模态的检测方法的优缺点总结。

4　未来展望

基于空间域的检测方法直观且较为通用，但抗扰动能力差；而基于频率域的检测方法对扰动具有一定鲁棒性，但在面对未知的伪造方法时检测能力有所下降；基于时空不一致的检测方法适用性广，但容易受到噪声等扰动的影响；基于生物特征的检测方法则另辟蹊径，直接检测视频中人的生物特征，因而泛化性显著提升，但对数据集要求较高；相比之下，基于多模态的检测方法结合多模态特征，鲁棒性和泛化性都较为良好，但预处理较为复杂，且对数据集要求较高。

表 5　基于多模态的检测方法优缺点总结

Table 5　Summary of the advantages and disadvantages of multimodal feature-based detection methods

方法	优点	缺点	实验数据集	实验结果/%	跨数据集	实验结果/%
Cai等人(2022)	端到端训练	只针对特定伪造情境	LAV-DF	AUC = 99.00	–	–
Shahzad等人(2022)	对噪声和干扰更鲁棒	依赖唇部与音频	FakeAVCeleb	ACC = 94.00	–	–
AVoiD-DF(Yang等,2023b)	充分捕捉跨模态关系	训练过程复杂	FakeAVCeleb	AUC = 89.20	DFDC	AUC = 80.60
AVFF(Oorloff等,2024)	高精度与泛化能力	不支持单一模态视频	FakeAVCeleb	AUC = 99.10	DFDC	AUC = 86.20
Yin等人(2024)	精度高且鲁棒	参数量大且设置复杂	FakeAVCeleb	AUC = 99.97	LAV-DF	AUC = 63.80
Song等人(2024)	能检测多类伪造视频	计算复杂度高	自建库	AUC = 95.70	–	–

注:“–”表示未进行跨库测试。

综上所述,深度伪造检测目前面临的主要瓶颈是如何在保证模型检测精度的同时兼顾模型的泛化能力,并使模型对各种扰动具有鲁棒性。

由于深度伪造检测常依赖于人脸定位、PPG信号提取等基础任务,因此迁移学习成为深度伪造检测领域的重要趋势。例如,许多研究使用Vision Transformer(Zhang等,2022)与Swin Transformer(Xu等,2024)作为底层预训练模型,之后在深度伪造数据集微调以提高分类性能。为了减少对标注数据的依赖,自监督学习近些年得以应用于深度伪造检测(Oorloff等,2024)。这种学习方法迫使模型学习并非因数据集不同而不同的深层特征,因而能提升模型的泛化性。除此之外,多实例学习(Li等,2020c)、多任务学习(Yu等,2024)等也在深度伪造检测领域得到应用。针对现有问题及趋势,该领域未来可能的研究方向如下:

1)提高检测模型的泛化性。目前的检测方法普遍存在泛化能力不足的问题,未来的研究应致力于提高检测模型的泛化性,以推动伪造检测的实际应用。可能的方法有:寻找共性且难以消除的不一致特征,推动模型学习更本质的特征;采用增量学习(Khan和Dai,2021)使模型能快速适应新伪造技术;采用自监督(Khormali和Yuan,2024)或无监督(Zhuang等,2022;Fung等,2021)训练方法等。

2)研究鲁棒性更强的检测方法。应从模型设计层面提高模型的鲁棒性,例如,采用模块化设计,使模型在面对新型伪造时能快速调整;选择更具辨别力的特征,减少对噪声和干扰的依赖。同时,在模型的训练中应该对数据进行旋转、缩放等扰动以探讨模型的鲁棒性,还可以通过生成对抗性样本(Aneja等,2022)以探讨其健壮性。

3)应用迁移学习(Ranjan等,2020)与跨域学习(Song等,2022b)。通过迁移学习,利用在大规模数据集上预训练的多模态模型,快速适应特定领域的深度伪造检测。此外,可结合跨域学习的方法,将模型在不同的应用场景或数据集上进行训练,使其具备跨领域的泛化能力。

4)创新学习方法。目前深度伪造检测领域的数据集质量参差不齐且特征差异大,可通过元学习(Sun等,2021a)、孪生网络以及集成学习(Hashmi等,2022)等方法克服数据集缺点、提升效果。

5)实时检测研究。传统的深度伪造检测方法通常需要大量计算资源,难以实现实时处理和在移动设备上的部署。未来应设计自适应检测系统或轻量级、低延迟的检测系统。

6)构建高质量数据集与全面的评价指标。当前,针对扩散模型生成人脸的检测缺乏通用数据集,且现有数据集大多在内容质量上显著低于最新的伪造技术水平。应基于当前先进的伪造技术构建高质量、精细化且具有通用性的数据集,并尽可能向数据集引入多模态信息。在评价指标方面,应根据推理时间、消耗资源等指标,构建更全面的评估框架。

7)主动防御(瞿左珉 等,2024)与内容溯源研究(刘安安 等,2024)。检测只是被动防御,未来应加强对主动防御的研究,如向视频添加水印(宋佳维等,2024)、加密签名或对抗性扰动(Aneja等,2022;Wang等,2022b)等,将伪造消除于根源或传播途中。此外,通过图像取证技术、区块链溯源等手段追踪伪造内容的生成工具、初始发布者和传播途径,为司法鉴定提供技术支持,也是一个重要的研究方向。

5 结 语

人脸深度伪造技术在影视创作、虚拟教学等领域应用广泛，但也常被不法分子滥用于制作虚假新闻或实施诈骗。该技术的发展大致经历了3个阶段：第1阶段主要基于传统图像处理，操作繁杂且真实度较低；第2阶段主要基于生成对抗网络，该技术的引入使伪造内容的逼真度显著提高；自2022年以来，随着扩散模型的兴起，人脸深度伪造迈入第3阶段，其生成内容质量更高、细节更丰富，但由于技术的前沿性，目前针对扩散模型生成人脸的研究较少，相关数据集较为稀缺，亟需更多探索与突破。

人脸深度伪造检测技术的发展始终紧随伪造技术。本文总结了人脸深度伪造检测领域的最新研究进展，与以往的综述相比，总结了针对扩散模型生成的图像/视频的检测方法及相关数据集，并整理了基于多模态的检测方法。具体而言，首先介绍了人脸深度伪造检测领域的常用数据集与评价指标。然后，根据各检测方法的特征选择的不同，对图像级与视频级伪造人脸检测方法进行了细致分类与总结。最后，探讨了人脸深度伪造检测研究的瓶颈与未来研究方向。

参考文献（References）

Afchar D, Nozick V, Yamagishi J and Echizen I. 2018. MesoNet: a compact facial video forgery detection network//Proceedings of the 10th IEEE International Workshop on Information Forensics and Security (WIFS). Hong Kong, China: IEEE: 1-7 [DOI: 10.1109/WIFS.2018.8630761]

Agarwal A, Agarwal A, Sinha S, Vatsa M and Singh R. 2021. MD-CSDNetwork: multi-domain cross stitched network for deepfake detection//Proceedings of the 16th IEEE International Conference on Automatic Face and Gesture Recognition. Jodhpur, India: IEEE: 1-8 [DOI: 10.1109/FG52635.2021.9666937]

Aghasanli A, Kangin D and Angelov P. 2023. Interpretable-through-prototypes deepfake detection for diffusion models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: 467-474 [DOI: 10.1109/ICCVW60793.2023.00053]

Amoroso R, Morelli D, Cornia M, Baraldi L, Del Bimbo A and Cucchiara R. 2024. Parents and children: distinguishing multimodal deepfakes from natural images. ACM Transactions on Multimedia Computing, Communications and Applications, 21(1): 1-23 [DOI: 10.1145/3665497]

Aneja S and Nießner M. 2020. Generalized zero and few-shot transfer for facial forgery detection [EB/OL]. [2024-11-07]. https://arxiv.org/pdf/2006.11863.pdf

Aneja S, Markhasin L and Nießner M. 2022. TAFIM: targeted adversarial attacks against facial image manipulations//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 58-75 [DOI: 10.1007/978-3-031-19781-9_4]

Binh L M and Woo S. 2022. ADD: frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images//Proceedings of the 36th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI: 122-130 [DOI: 10.1609/aaai.v36i1.19886]

Cai Z X, Stefanov K, Dhall A and Hayat M. 2022. Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization//Proceedings of 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA). Sydney, Australia: IEEE: 1-10 [DOI: 10.1109/DICTA56598.2022.10034605]

Cao J Y, Ma C, Yao T P, Chen S, Ding S H and Yang X K. 2022. End-to-end reconstruction-classification learning for face forgery detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 4103-4112 [DOI: 10.1109/CVPR52688.2022.00408]

Cao S H, Liu X H, Mao X Q and Zou Q. 2022. A review of human face forgery and forgery-detection technologies. Journal of Image and Graphics, 27(4): 1023-1038 (曹申豪, 刘晓辉, 毛秀青, 邹勤. 2022. 人脸伪造及检测技术综述. 中国图象图形学报, 27(4): 1023-1038) [DOI: 10.11834/jig.200466]

Chandra K A, Aashutosh A V, Das S and Das A. 2024. Latent flow diffusion for deepfake video generation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 3781-3790

Chen H, Lin Y Z and Li B. 2023. Exposing face forgery clues via retinex-based image enhancement//Proceedings of the 16th Asian Conference on Computer Vision. Macao, China: Springer: 20-34 [DOI: 10.1007/978-3-031-26316-3_2]

Chen L, Zhang Y, Song Y B, Liu L Q and Wang J. 2022. Self-supervised learning of adversarial example: towards good generalizations for deepfake detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 18689-18698 [DOI: 10.1109/CVPR52688.2022.01815]

Chen P, Liu J, Liang T, Zhou G Z, Gao H C, Dai J and Han J Z. 2020. FSSPOTTER: spotting face-swapped video by spatial and temporal clues//Proceedings of 2020 IEEE International Conference on Multimedia and Expo (ICME). London, UK: IEEE: 1-6 [DOI: 10.1109/ICME46284.2020.9102914]

Chen S, Yao T P, Chen Y, Ding S H, Li J L and Ji R R. 2021. Local relation learning for face forgery detection//Proceedings of the 35th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI: 1081-1088 [DOI: 10.1609/aaai.v35i2.16193]

Chen W X and McDuff D. 2018. DeepPhys: video-based physiological measurement using convolutional attention networks//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 356-373 [DOI: 10.1007/978-3-030-01216-8_22]

Cheng H, Guo Y Y, Wang T Y, Nie L Q and Kankanhalli M. 2024. Diffusion facial forgery detection//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM: 5939-5948 [DOI: 10.1145/3664647.3680797]

Ciftci U A, Demir I and Yin L. 2020. FakeCatcher: detection of synthetic portrait videos using biological signals. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2(1): 1-26 [DOI: 10.1109/TPAMI.2020.3009287]

Dai Y S, Fei J W, Xia Z H, Liu J N and Wong J. 2023. Local similarity anomaly for general face forgery detection. Journal of Image and Graphics, 28(11): 3453-3470 (戴昀书, 费建伟, 夏志华, 刘家男, 翁健. 2023. 局部相似度异常的强泛化性伪造人脸检测. 中国图象图形学报, 28(11): 3453-3470) [DOI: 10.11834/jig.221006]

Dolhansky B, Bitton J, Pflaum B, Lu J K, Howes R, Wang M L and Ferrer C C. 2020. The deepfake detection challenge (DFDC) dataset [EB/OL]. [2024-11-07]. https://arxiv.org/pdf/2006.07397.pdf

Durall R, Keuper M, Pfreundt F J and Keuper J. 2019. Unmasking deepfakes with simple features [EB/OL]. [2024-11-07]. https://arxiv.org/pdf/1911.00686.pdf

Feng C B, Liu C X, Wang Y Y and Zhou Q D. 2024. Face forgery detection with image patch comparison and residual map estimation. Journal of Image and Graphics, 29(2): 457-467 (冯才博, 刘春晓, 王昱烨, 周其当. 2024. 结合图像块比较与残差图估计的人脸伪造检测. 中国图象图形学报, 29(2): 457-467) [DOI: 10.11834/jig.230149]

Feng C, Chen Z Y and Owens A. 2023. Self-supervised video forensics by audio-visual anomaly detection//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 10491-10503 [DOI: 10.1109/CVPR52729.2023.01011]

Frank J, Eisenhofer T, Schönherr L, Fischer A, Kolossa D and Holz T. 2020. Leveraging frequency analysis for deep fake image recognition//Proceedings of the 37th International Conference on Machine Learning. [s.l.]: ACM: #304

Fung S, Lu X Q, Zhang C and Li C S. 2021. DeepfakeUCL: deepfake detection via unsupervised contrastive learning//Proceedings of 2021 International Joint Conference on Neural Networks (IJCNN). Shenzhen, China: IEEE: 1-8 [DOI: 10.1109/IJCNN52387.2021.9534089]

Gandhi A and Jain S. 2020. Adversarial perturbations fool deepfake detectors//Proceedings of 2020 International Joint Conference on Neural Networks (IJCNN). Glasgow, UK: IEEE: 1-8 [DOI: 10.1109/IJCNN48605.2020.9207034]

Gu Q Q, Chen S, Yao T P, Chen Y, Ding S H and Yi R. 2022a. Exploiting fine-grained face forgery clues via progressive enhancement learning//Proceedings of the 36th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI: 735-743 [DOI: 10.1609/aaai.v36i1.19954]

Gu Z H, Chen Y, Yao T P, Ding S H, Li J L and Ma L Z. 2022b. Delving into the local: dynamic inconsistency learning for deepfake video detection//Proceedings of the 36th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI: 744-752 [DOI: 10.1609/aaai.v36i1.19955]

Güera D and Delp E J. 2018. Deepfake video detection using recurrent neural networks//Proceedings of the 15th IEEE international conference on advanced video and signal based surveillance (AVSS). Auckland, New Zealand: IEEE: 1-6 [DOI: 10.1109/AVSS.2018.8639163]

Haliassos A, Mira R, Petridis S and Pantic M. 2022. Leveraging real talking faces via self-supervision for robust forgery detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 14930-14942 [DOI: 10.1109/CVPR52688.2022.01453]

Haliassos A, Vougioukas K, Petridis S and Pantic M. 2021. Lips don't lie: a generalisable and robust approach to face forgery detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 5037-5047 [DOI: 10.1109/CVPR46437.2021.00500]

Hashmi A, Shahzad S A, Ahmad W, Lin C W, Tsao Y and Wang H M. 2022. Multimodal forgery detection using ensemble learning//Proceedings of 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Chiang Mai, Thailand: IEEE: 1524-1532 [DOI: 10.23919/APSIPAASC55919.2022.9980255]

Heo Y J, Yeo W H and Kim B G. 2023. DeepFake detection algorithm based on improved vision transformer. Applied Intelligence, 53(7): 7512-7527 [DOI: 10.1007/s10489-022-03867-9]

Hernandez-Ortega J, Tolosana R, Fiérrez J and Morales A. 2020. DeepFakesOn-Phys: Deepfakes detection based on heart rate estimation//Proceedings of the Workshop on Artificial Intelligence Safety 2021. [s.l.]: CEUR

Hooda A, Mangaokar N, Feng R, Fawaz K, Jha S and Prakash A. 2024. D4: detection of adversarial diffusion deepfakes using disjoint ensembles//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 3800-3810 [DOI: 10.1109/WACV57701.2024.00377]

Hu S, Li Y Z and Lyu S W. 2021. Exposing GAN-generated faces using inconsistent corneal specular highlights//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Process-

ing (ICASSP). Toronto, Canada: IEEE: 2500-2504 [DOI: 10.1109/ICASSP39728.2021.9414582]

Jiang Y M, Huang Z Q, Pan X G, Loy C C and Liu Z W. 2021. Talk-to-edit: fine-grained facial editing via dialog//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 13779-13788 [DOI: 10.1109/ICCV48922.2021.01354]

Karras T, Laine S and Aila T. 2021. A style-based generator architecture for generative adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43 (12): 4217-4228 [DOI: 10.1109/TPAMI.2020.2970919]

Khalid H, Tariq S, Kim M and Woo S S. 2021. FakeAVCeleb: a novel audio-video multimodal deepfake dataset//Proceedings of the 35th Conference on Neural Information Processing Systems Track on Datasets and Benchmarks Track (Round 2. NeurIPS 2021): [s.l.]: [s.n.]

Khan S A and Dai H. 2021. Video transformer for deepfake detection with incremental learning//Proceedings of the 29th ACM International Conference on Multimedia. [s.l.]: ACM: 1821-1828 [DOI: 10.1145/3474085.3475332]

Khormali A and Yuan J S. 2024. Self-supervised graph transformer for deepfake detection. IEEE Access, 12: 58114-58127 [DOI: 10.1109/ access.2024.3392512]

Korshunov P and Marcel S. 2018. DeepFakes: a new threat to face recognition? assessment and detection [EB/OL]. [2024-11-07]. https://arxiv.org/pdf/1812.08685.pdf

Korshunova I, Shi W Z, Dambre J and Theis L. 2017. Fast face-swap using convolutional neural networks//Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 3697-3705 [DOI: 10.1109/ICCV.2017.397]

Lee S, Tariq S, Kim J and Woo S S. 2021. TAR: generalized forensic framework to detect deepfakes using weakly supervised learning//Proceedings of the 36th IFIP TC 11 International Conference on ICT Systems Security and Privacy Protection. Oslo, Norway: Springer: 351-366 [DOI: 10.1007/978-3-030-78120-0_23]

Lewis J K, Toubal I E, Chen H L, Sandesera V, Lomnitz M, Hampel-Arias Z, Prasad C and Palaniappan K. 2020. Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning//Proceedings of 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR). Washington, USA: IEEE: 1-9 [DOI: 10.1109/AIPR50011.2020.9425167]

Li H D, Li B, Tan S Q and Huang J W. 2020a. Identification of deep network generated images using disparities in color components. Signal Processing, 174: #107616 [DOI: 10.1016/j.sigpro.2020.107616]

Li J M, Xie H T, Li J H, Wang Z Y and Zhang Y D. 2021. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 6454-6463 [DOI: 10.1109/CVPR46437.2021.00639]

Li J M, Xie H T, Yu L Y and Zhang Y D. 2022. Wavelet-enhanced weakly supervised local feature learning for face forgery detection//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: ACM: 1299-1308 [DOI: 10.1145/3503161.3547832]

Li L Z, Bao J M, Zhang T, Yang H, Chen D, Wen F and Guo B N. 2020b. Face X-ray for more general face forgery detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 5000-5009 [DOI: 10.1109/CVPR42600.2020.00505]

Li X D, Lang Y N, Chen Y F, Mao X F, He Y, Wang S H, Xue H and Lu Q. 2020c. Sharp multiple instance learning for deepfake video detection//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 1864-1872 [DOI: 10.1145/3394171.3414034]

Li Y Z and Lyu S W. 2018. Exposing deepfake videos by detecting face warping artifacts//Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, USA: IEEE: 46-52

Li Y Z, Chang M C and Lyu S W. 2018. In ictu oculi: exposing ai created fake videos by detecting eye blinking//Proceedings of 2018 IEEE International Workshop on Information Forensics and Security (WIFS). Hong Kong, China: IEEE: 1-7 [DOI: 10.1109/WIFS.2018.8630787]

Li Y Z, Yang X, Sun P, Qi H G and Lyu S W. 2020d. Celeb-DF: a large-scale challenging dataset for deepfake forensics//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 3204-3213 [DOI: 10.1109/CVPR42600.2020.00327]

Li Y, Bian S, Wang C T and Lu W. 2023. CNN and Transformer-coordinated deepfake detection. Journal of Image and Graphics, 28(3): 804-819 (李颖, 边山, 王春桃, 卢伟. 2023. CNN结合Transformer的深度伪造高效检测. 中国图象图形学报, 28(3): 804-819) [DOI: 10.11834/jig.220519]

Liu A A, Su Y T, Wang L J, Li B, Qian Z X, Zhang W M, Zhou L N, Zhang X P, Zhang Y D, Huang J W and Yu N H. 2024. Review on the progress of the AIGC visual content generation and traceability. Journal of Image and Graphics, 29(6): 1535-1554 (刘安安, 苏育挺, 王岚君, 李斌, 钱振兴, 张卫明, 周琳娜, 张新鹏, 张勇东, 黄继武, 俞能海. 2024. AIGC视觉内容生成与溯源研究进展. 中国图象图形学报, 29(6): 1535-1554) [DOI: 10.11834/jig.240003]

Liu B P, Liu B, Ding M and Zhu T Q. 2024. Detection of diffusion model-generated faces by assessing smoothness and noise tolerance//The 2024 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB). Toronto, Canada: IEEE: 1-6 [DOI: 10.1109/BMSB62888.2024.10608232]

Lorenz P, Durall R L and Keuper J. 2023. Detecting images generated by deep diffusion models using their local intrinsic dimensionality// Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 448-459 [DOI: 10.1109/ICCVW60793.2023.00051]

Lu W, Sun W and Lu H T. 2009. Robust watermarking based on DWT and nonnegative matrix factorization. Computers and Electrical Engineering, 35(1): 183-188 [DOI: 10.1016/j.compeleceng.2008.09.004]

Mandelli S, Bonettini N, Bestagini P and Tubaro S. 2022. Detecting GAN-generated images by orthogonal training of multiple CNNs// Proceedings of 2022 IEEE International Conference on Image Processing (ICIP). Bordeaux, France: IEEE: 3091-3095 [DOI: 10.1109/ICIP46576.2022.9897310]

Mao M Y and Yang J. 2021. Exposing deepfake with pixel-wise AR and PPG correlation from faint signals [EB/OL]. [2024-11-07]. https://arxiv.org/pdf/2110.15561.pdf

Matern F, Riess C and Stamminger M. 2019. Exploiting visual artifacts to expose deepfakes and face manipulations//Proceedings of 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). Waikoloa, USA: IEEE: 83-92 [DOI: 10.1109/WACVW.2019.00020]

Mehta D, Mehta A and Narang P. 2024. LDFaceNet: latent diffusion-based network for high-fidelity deepfake generation//Proceedings of the 27th International Conference on Pattern Recognition. Kolkata, India: Springer: 386-400 [DOI: 10.1007/978-3-031-78389-0_26]

Miao Q M, Kang S, Marsella S, DiPaola S, Wang C and Shapiro A. 2022. Study of detecting behavioral signatures within DeepFake videos [EB/OL]. [2024-11-07]. https://arxiv.org/pdf/2208.03561.pdf

Nataraj L, Mohammed T M, Chandrasekaran S, Flenner A, Bappy J H, Roy-Chowdhury A K and Manjunath B S 2019. Detecting GAN generated fake images using co-occurrence matrices. Electronic Imaging, 31 (5): 1-7 [DOI: 10.2352/issn. 2470-1173.2019.5.mwsf-532]

Nguyen H H, Yamagishi J and Echizen I. 2019. Capsule-forensics: using capsule networks to detect forged images and videos//Proceedings of the 44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE: 2307-2311 [DOI: 10.1109/ICASSP.2019.8682602]

Oorloff T, Koppisetti S, Bonettini N, Solanki D, Colman B, Yacoob Y, Shahriyari A and Bharaj G. 2024. AVFF: audio-visual feature fusion for video deepfake detection//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 27092-27102 [DOI: 10.1109/CVPR52733.2024.02559]

Papa L, Faiella L, Corvitto L, Maiano L and Amerini I. 2023. On the use of stable diffusion for creating realistic faces: from generation to detection//Proceedings of the 11th International Workshop on Biometrics and Forensics (IWBF). Barcelona, Spain: IEEE: 1-6 [DOI: 10.1109/IWBF57495.2023.10156981]

Qian Y Y, Yin G J, Sheng L, Chen Z X and Shao J. 2020. Thinking in frequency: face forgery detection by mining frequency-aware clues// Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 86-103 [DOI: 10.1007/978-3-030-58610-2_6]

Qu Z M, Yin Q L, Sheng Z Q, Wu J Y, Zhang B L, Yu S R and Lu W. 2024. Overview of deepfake proactive defense techniques. Journal of Image and Graphics, 29(2): 318-342 (瞿左珉, 殷琪林, 盛紫琦, 吴俊彦, 张博林, 余尚戎, 卢伟. 2024. 人脸深度伪造主动防御技术综述. 中国图象图形学报, 29(2): 318-342) [DOI: 10.11834/jig.230128]

Ranjan P, Patil S and Kazi F. 2020. Improved generalizability of deepfakes detection using transfer learning based CNN framework//Proceedings of the 3rd International Conference on Information and Computer Technologies (ICICT). San Jose, USA: IEEE: 86-90 [DOI: 10.1109/ICICT50521.2020.00021]

Ren J Q, Qin J P, Ma Q L and Cao Y. 2024. FastFaceCLIP: a lightweight text-driven high-quality face image manipulation. IET Computer Vision, 18(7): 950-967 [DOI: 10.1049/cvi2.12295]

Ricker J, Damm S, Holz T and Fischer A. 2022. Towards the detection of diffusion model deepfakes [EB/OL]. [2024-11-07]. https://arxiv.org/pdf/2210.14571.pdf

Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J and Niessner M. 2019. FaceForensics++: learning to detect manipulated facial images//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 1-11 [DOI: 10.1109/ICCV.2019.00009]

Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J and Nießner M. 2018. FaceForensics: a large-scale video dataset for forgery detection in human faces [EB/OL]. [2024-11-07]. https://arxiv.org/pdf/1803.09179.pdf

Sabir E, Cheng J X, Jaiswal A, AbdAlmageed W, Masi I and Natarajan P. 2019. Recurrent convolutional strategies for face manipulation detection in videos//Proceedings of 2019 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, USA: IEEE: 80-87

Saif S, Tehseen S and Ali S S. 2024. Fake news or real? detecting deepfake videos using geometric facial structure and graph neural network. Technological Forecasting and Social Change, 205: #123471 [DOI: 10.1016/j.techfore.2024.123471]

Shahzad S A, Hashmi A, Khan S, Peng Y T, Tsao Y and Wang H M. 2022. Lip-sync matters: a novel multimodal forgery detector//Proceedings of 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Chiang Mai, Thailand: IEEE: 1885-1892 [DOI: 10.23919/APSIPAASC55919.2022.9980296]

Song H K, Woo S H, Lee J, Yang S, Cho H, Lee Y, Choi D and Kim K W. 2022a. Talking face generation with multilingual TTs//Pro-

ceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 21393-21398 [DOI: 10.1109/CVPR52688.2022.02074]

Song J W, Liu C X and Zhang X Y. 2024. LDH: least dependent hiding for screen-shooting resilient watermarking. Journal of Image and Graphics, 29(2): 408-418 (宋佳维, 刘春晓, 张心怡. 2024. 最小依赖隐藏的屏摄鲁棒水印方法. 中国图象图形学报, 29(2): 408-418) [DOI: 10.11834/jig.220811]

Song L C, Fang Z, Li X D, Dong X Y, Jin Z C, Chen Y F and Lyu S W. 2022b. Adaptive face forgery detection in cross domain//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 467-484 [DOI: 10.1007/978-3-031-19830-4_27]

Song X F, Guo X, Zhang J C, Li Q R, Bai L, Liu X M, Zhai G T and Liu X H. 2024. On learning multi-modal forgery representation for diffusion generated video detection//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS

Stefanov K, Paliwal B and Dhall A. 2022. Visual representations of physiological signals for fake video detection [EB/OL]. [2024-11-07]. https://arxiv.org/pdf/2207.08380.pdf

Sun K, Liu H, Ye Q X, Gao Y, Liu J Z, Shao L and Ji R R. 2021a. Domain general face forgery detection by learning to weight//Proceedings of the 35th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI: 2638-2646 [DOI: 10.1609/aaai.v35i3.16367]

Sun K, Yao T P, Chen S, Ding S H, Li J L and Ji R R. 2022. Dual contrastive learning for general face forgery detection//Proceedings of the 36th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI: 2316-2324 [DOI: 10.1609/aaai.v36i2.20130]

Sun Z K, Han Y J, Hua Z Y, Ruan N and Jia W J. 2021b. Improving the efficiency and robustness of deepfakes detection through precise geometric features//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 3608-3617 [DOI: 10.1109/CVPR46437.2021.00361]

Țânțaru D C, Oneață E and Oneață D. 2024. Weakly-supervised deepfake localization in diffusion-generated images//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 6246-6256 [DOI: 10.1109/WACV57701.2024.00614]

Thies J, Zollhöfer M and Nießner M. 2019a. Deferred neural rendering: image synthesis using neural textures. ACM Transactions on Graphics, 38(4): #66 [DOI: 10.1145/3306346.3323035]

Thies J, Zollhöfer M, Stamminger M, Theobalt C and Nießner M. 2019b. Face2Face: real-time face capture and reenactment of RGB videos. Communications of the ACM, 62(1): 96-104 [DOI: 10.1145/3292039]

Wang C R and Deng W H. 2021. Representative forgery mining for fake face detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14918-14927 [DOI: 10.1109/CVPR46437.2021.01468]

Wang J K, Wu Z X, Ouyang W H, Han X T, Chen J J, Jiang Y G and Li S N. 2022a. M2TR: multi-modal multi-scale transformers for deepfake detection//Proceedings of 2022 International Conference on Multimedia Retrieval. Newark, USA: ACM: 615-623 [DOI: 10.1145/3512527.3531415]

Wang R, Huang Z H, Chen Z K, Liu L, Chen J and Wang L N. 2022b. Anti-forgery: towards a stealthy and robust deepfake disruption attack via adversarial perceptual-aware perturbations//Proceedings of the 31st International Joint Conference on Artificial Intelligence. Vienna, Austria: IJCAI: 761-767

Wang Z D, Bao J M, Zhou W G, Wang W L, Hu H Z, Chen H and Li H Q. 2023a. DIRE for diffusion-generated image detection//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 22388-22398 [DOI: 10.1109/ICCV51070.2023.02051]

Wang Z J, Montoya E, Munechika D, Yang H Y, Hoover B and Chau D H. 2023b. DiffusionDB: a large-scale prompt gallery dataset for text-to-image generative models//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada: ACL: 893-911

Wodajo D, Atnafu S and Akhtar Z. 2023. Deepfake video detection using generative convolutional vision transformer [EB/OL]. [2024-11-07]. https://arxiv.org/pdf/2307.07036v1.pdf

Xia W H, Yang Y J, Xue J H and Wu B Y. 2021. TediGAN: text-guided diverse face image generation and manipulation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 2256-2265 [DOI: 10.1109/CVPR46437.2021.00229]

Xu C, Zhang J N, Hua M, He Q, Yi Z L and Liu Y. 2022a. Region-aware face swapping//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 7622-7631 [DOI: 10.1109/CVPR52688.2022.00748]

Xu Y B, Yin Y Q, Jiang L M, Wu Q Y, Zheng C Y, Loy C C, Dai B and Wu W. 2022b. TransEditor: transformer-based dual-space GAN for highly controllable facial editing//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 7673-7682 [DOI: 10.1109/CVPR52688.2022.00753]

Xu Y T, Liang J, Sheng L J and Zhang X Y. 2024. Learning spatiotemporal inconsistency via thumbnail layout for face deepfake detection. International Journal of Computer Vision, 132(12): 5663-5680 [DOI: 10.1007/s11263-024-02054-2]

Yang P N, Huang H B, Wang Z Y, Yu A J and He R. 2023a. Confidence-calibrated face image forgery detection with contrastive representation distillation//Proceedings of the 16th Asian Conference on Computer Vision. Macao, China: Springer: 3-19 [DOI: 10.1007/978-3-031-26316-3_1]

Yang W Y, Zhou X Y, Chen Z K, Guo B F, Ba Z J, Xia Z H, Cao X C and Ren K. 2023b. AVoiD-DF: audio-visual joint learning for detecting deepfake. IEEE Transactions on Information Forensics and Security, 18: 2015-2029 [DOI: 10.1109/tifs.2023.3262148]

Yin Q L, Lu W, Cao X C, Luo X Y, Zhou Y C and Huang J W. 2024. Fine-grained multimodal deepfake classification via heterogeneous graphs. International Journal of Computer Vision, 132(11): 5255-5269 [DOI: 10.1007/s11263-024-02128-1]

Yu Z T, Cai R Z, Li Z, Yang W H, Shi J G and Kot A C. 2024. Benchmarking joint face spoofing and forgery detection with visual and physiological cues. IEEE Transactions on Dependable and Secure Computing, 21 (5): 4327-4342 [DOI: 10.1109/TDSC. 2024. 3352049]

Zhang C X, Zhao Y F, Huang Y F, Zeng M, Ni S F, Budagavi M and Guo X H. 2021. FACIAL: synthesizing dynamic talking face with implicit attribute learning//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 3847-3856 [DOI: 10.1109/ICCV48922.2021.00384]

Zhang D C, Lin F Z, Hua Y Y, Wang P J, Zeng D and Ge S M. 2022. Deepfake video detection with spatiotemporal dropout transformer// Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: ACM: 5833-5841 [DOI: 10.1145/3503161. 3547913]

Zhang X, Karaman S and Chang S F. 2019. Detecting and simulating artifacts in GAN fake images//Proceedings of 2019 IEEE International Workshop on Information Forensics and Security (WIFS). Delft, the Netherlands: IEEE: 1-6 [DOI: 10.1109/wifs47025. 2019.9035107]

Zhao H Q, Wei T Y, Zhou W B, Zhang W M, Chen D D and Yu N H. 2021. Multi-attentional deepfake detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 2185-2194 [DOI: 10.1109/ CVPR46437.2021.00222]

Zheng Y L, Bao J M, Chen D, Zeng M and Wen F. 2021. Exploring temporal coherence for more general video face forgery detection// Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 15024-15034 [DOI: 10. 1109/ICCV48922.2021.01477]

Zhou P, Han X T, Morariu V I and Davis L S. 2017. Two-stream neural networks for tampered face detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu, USA: IEEE: 1831-1839 [DOI: 10. 1109/CVPRW.2017.229]

Zhu X T, Tang Y Q and Geng P Z. 2021. Detection algorithm of tamper and deepfake image based on feature fusion. Netinfo Security, 21(8): 70-81 (朱新同, 唐云祁, 耿鹏志. 2021. 基于特征融合的篡改与深度伪造图像检测算法. 信息网络安全, 21(8): 70-81) [DOI: 10.3969/j.issn.1671-1122.2021.08.009]

Zhuang W Y, Chu Q, Tan Z T, Liu Q K, Yuan H J, Miao C T, Luo Z X and Yu N H. 2022. UIA-ViT: unsupervised inconsistency-aware method based on vision transformer for face forgery detection//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 391-407 [DOI: 10.1007/978-3-031-20065-6_23]

Zhuo W Q, Li D Z, Wang W and Dong J. 2023. Data-free model compression for light-weight DeepFake detection. Journal of Image and Graphics, 28(3): 820-835 (卓文琦, 李东泽, 王伟, 董晶. 2023. 面向轻量级深度伪造检测的无数据模型压缩. 中国图象图形学报, 28(3): 820-835) [DOI: 10.11834/jig.220559]

作者简介

姚文达,男,本科生,主要研究方向为伪造人脸检测。
E-mail: 220701240209@stu.nepu.edu.cn
赵娅,通信作者,女,副教授,主要研究方向为图像处理、机器学习和深度学习。E-mail:zhaoya@nepu.edu.cn
李盼池,男,教授,主要研究方向为图像处理、智能计算和深度学习。E-mail:lipanchi@vip.163.com
吴洪超,男,二级警长,主要研究方向为网络安全管理、网络安全保护和内容安全。E-mail:wuhongchao0802@163.com