E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP18;TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2026)01-0082-17

论文引用格式:Bei Y J, Lou H R, Gao K W, Song J, Wang R, Jin C H, Lei J, Song M L, Hu B D and Feng Z L. 2026. Large-scale Chinese data evaluation benchmark for face video forgery detection. Journal of Image and Graphics, 31(1):0082-0098(贝毅君, 娄恒瑞, 高克威, 宋杰, 王蕊, 金苍宏, 雷杰, 宋明黎, 胡秉德, 冯尊磊. 2026. 面向人脸视频防伪检测的大规模中文数据测评基准. 中国图象图形学报, 31(1):0082-0098)[DOI:10.11834/jig.250077]

面向人脸视频防伪检测的大规模中文数据测评基准

贝毅君[1],娄恒瑞[2],高克威[1],宋杰[1,2],王蕊[3],金苍宏[4],雷杰[5],宋明黎[2],胡秉德[2*],冯尊磊[1,2]

1. 浙江大学软件学院,宁波　315103; 2. 浙江大学计算机科学与技术学院,杭州　310007;
3. 中国科学院信息工程研究所,北京　100093; 4. 浙大城市学院计算机与计算科学学院,杭州　310015;
5. 浙江工业大学计算机科学与技术学院,杭州　310023

摘　要:**目的**　针对生成式人工智能(artificial intelligence generated content,AIGC)技术生成的高逼真伪造人脸视频对人类视觉感知的欺骗性问题,以及当前人脸防伪检测算法评估体系在中文数据层面有效性和应用性验证方面的空白,旨在构建面向中文场景的量化评估基准以推动防伪检测技术迭代发展。**方法**　提出面向大规模中文人脸伪造视频的CHN-DF(Chinese-deepfake)数据集,详细阐述数据采集、伪造样本生成及质量评估的全流程构建方法。通过多维度实验验证数据集复杂性,兼顾跨模态伪造技术覆盖、环境干扰因子完备性等复杂因素,并建立基于深度检测模型的系统性评测基准。**结果**　发布全球首个包含434 727样本的中文人脸视频防伪数据集,实验显示该数据集鉴别难度高,在16种包含SOTA(state-of-the-art)与主流防伪模型的测评中视觉与视听结合的准确率分别控制在85%与70%以下。构建的评测基准覆盖了视觉与听觉模态场景,在跨域泛化性测试中显示模型准确率性能波动平均幅度达19.6%,显著揭示现有算法的应用局限性。**结论**　构建的中文防伪评测基准有效填补领域空白,通过系统性实验阐明数据集特性与算法性能的关联机制,提出针对模型鲁棒性增强、跨模态泛化能力提升等关键发展方向,为面向中文场景的量化评估以及人脸视频防伪技术的实际部署提供数据支撑与实践指导。CHN-DF数据集在线发布地址为:https://doi.org/10.57760/sciencedb.j00240.00067和https://github.com/HengruiLou/CHN-DF。

关键词:深度伪造;人脸伪造视频;人脸防伪评测基准;中文数据集;多模态

Large-scale Chinese data evaluation benchmark for face video forgery detection

Bei Yijun[1], Lou Hengrui[2], Gao Kewei[1], Song Jie[1,2], Wang Rui[3], Jin Canghong[4], Lei Jie[5], Song Mingli[2], Hu Bingde[2*], Feng Zunlei[1,2]

1. College of Software Technology, Zhejiang University, Ningbo 315103, China; 2. College of Computer Science and Technology, Zhejiang University, Hangzhou 310007, China; 3. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China; 4. College of Computer and Computing Science, Hangzhou City University, Hangzhou 310015, China; 5. College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China

Abstract: **Objective** The rapid advancement of AI-generated content (AIGC) technologies has enabled the creation of

收稿日期:2025-02-28;**修回日期**:2025-06-04;**预印本日期**:2025-06-11

*通信作者:胡秉德　tonyhu@zju.edu.cn

基金项目:国家重点研发计划资助(2022YFB2703100)

Supported by:National Key R&D Program of China (2022YFB2703100)

hyper-realistic synthetic face videos, posing unprecedented challenges to human visual perception systems. While existing face anti-spoofing detection algorithms demonstrate promising performance on Western-centric datasets, their effectiveness and applicability remain unverified in Chinese-specific contexts due to the lack of standardized evaluation benchmarks. This study aims to establish a quantitative assessment framework tailored for Chinese scenarios, driving the iterative development of anti-spoofing technologies. **Method** Aiming to address this issue, the CHN-DF dataset, the first large-scale benchmark dedicated to Chinese face video forgery detection, is proposed. The dataset construction process involves several critical stages, including data collection, fake sample generation, and quality assessment. Data collection was conducted in complex environmental settings, such as varying lighting conditions and occlusions, ensuring the diverse and challenging nature of the dataset. Over 190 000 dialogue clips from 2 540 real Chinese identity videos were collected across multiple sources, ensuring rich dataset variability. Twelve advanced AIGC tools that incorporate mainstream and cutting-edge methods based on diffusion, generative adversarial networks(GANs), and hybrid architectures are employed to generate the forged samples. These tools produced 434 727 mixed fake samples, covering seven types of manipulation techniques, such as expression transfer, lip-sync tampering, and audiovisual dissonance. Aiming to ensure data quality and integrity, human perception and automated detection techniques were employed in a dual-layer quality control process. This approach included an assessment framework that considers visual and auditory modalities to create a comprehensive evaluation standard that reflects the challenges encountered by anti-spoofing models in real-world applications. A system-level evaluation benchmark based on deep learning detection models was developed to further enhance the effectiveness of the dataset. This benchmark spans a variety of deepfake detection models, including state-of-the-art (SOTA) and widely used algorithms, to test the robustness and generalization capabilities of these models across different environmental and contextual factors. The goal was to test the adaptability of the detection algorithms to the complexities presented by the CHN-DF dataset. **Result** The CHN-DF dataset, which represents the world's first Chinese language-based face video anti-spoofing dataset, contains a total of 434 727 samples. This dataset is designed to be diverse and complex, posing substantial challenges for current deepfake detection algorithms. In experimental evaluations using 16 different models, including SOTA and mainstream anti-spoofing models, the accuracy of detecting deepfake faces in this dataset was found to be lower than expected, thereby highlighting task complexity. Specifically, the accuracy for visual and audiovisual combined tests was below 85% and 70%, respectively. These results indicate the increased difficulty in accurately detecting deepfake faces in the CHN-DF dataset, demonstrating that current models are still far from perfect. Further analysis revealed the effectiveness of the evaluation benchmark in highlighting the limitations of existing algorithms. In cross-domain generalization tests, the performance of models fluctuated by an average of 19.6%, indicating substantial variations in model accuracy across different scenarios. This variability emphasizes the need for robust and adaptable algorithms that can handle diverse environmental conditions and complex manipulations in the CHN-DF dataset. Moreover, the evaluation benchmark demonstrated the need for integrating visual and auditory modalities for face video anti-spoofing detection. In numerous cases, models that relied solely on visual features failed to accurately detect fake videos, highlighting the importance of incorporating multimodal features for improved detection accuracy. **Conclusion** The CHN-DF dataset and its associated evaluation benchmark effectively fill a critical gap in the field of face video anti-spoofing detection for Chinese language scenarios. Using systematic experimentation, this research clarifies the relationship between dataset characteristics and algorithm performance, elucidating the underlying mechanisms that contribute to the challenges faced by current models. Based on these findings, a roadmap for future improvements in the field is proposed, focusing on key directions such as enhancing model robustness and improving cross-modal generalization capabilities. These insights are crucial for the practical deployment of face video anti-spoofing detection technologies, offering theoretical support and practical guidance for future development. The CHN-DF dataset and evaluation protocol have been made publicly available on GitHub at https://github.com/HengruiLou/CHN-DF, serving as a valuable resource for researchers and practitioners in the field. This dataset serves not only as a technical reference but also as a catalyst for global collaboration in the fight against sophisticated, culturally specific deepfake threats. Future work will extend the dataset to include minority dialects. The dataset is linked at https://doi.org/10.57760/sciencedb.j00240.00067 and https://github.com/HengruiLou/CHN-DF.

Key words: deepfake; deepfake face videos; face forgery evaluation benchmark; Chinese dataset; multimodal

0 引 言

生成式人工智能(artificial intelligence generated content, AIGC)作为应对数字智能时代挑战的核心技术,通过语义理解与跨模态生成能力,已在艺术创作、场景编辑以及数字人生成等多领域展现出显著的应用价值(Han等,2023)。然而,“眼见为实”的理念往往深植人心,AIGC发展的同时也使得人脸视频伪造产物变得愈加逼真,信息真实性验证逼近失效(Zhao等,2025),导致当今社会面临严峻的安全威胁。对政治人物的深度伪造视频通过形象抹黑和内容篡改,可能影响国家政治制度甚至引发国际战争危机。此外,社交身份的伪造导致各类诈骗现象不断增多。一些不法分子利用深度伪造技术塑造虚假的个人形象,在聊天室中通过面孔和声音模拟与“同龄”儿童进行数字对话,以获取未成年人的信任,对他们的安全构成威胁。

为了应对人脸视频深度伪造技术的滥用和潜在危害,人脸视频防伪检测技术的研发与落地变得尤为重要,这些技术的落地依托于高质量的人脸视频防伪数据集。然而目前领域内数据集存在显著缺陷:1)在数据公平性上,人脸视频防伪检测数据集主要为欧美面孔,缺乏亚洲人脸视频数据样本,面向人脸视频防伪检测的中文数据仍是空白。2)在数据丰富性上,人脸视频防伪数据集落后于日新月异的现实AIGC生成场景,存在伪造模态单一、伪造方法老旧以及数据样本数量不足的共性问题。

为了应对上述问题,本文构建了首个面向人脸视频防伪检测的大规模中文数据集(Chinese-deepfake, CHN-DF),旨在构建面向中文场景的量化评估基准以推动人脸防伪检测技术迭代发展,其主要贡献如下:1)数据公平性。在视觉模态方面发布了针对中国面孔的人脸视频防伪数据集,填补了领域空白;在听觉模态方面构建了首个中文普通话语态的人脸防伪数据集,有助于推动普通话场景语音防伪技术发展。2)数据丰富性。CHN-DF数据量达到434 727个样本,是目前数据规模最大的人脸视频防伪数据集。伪造方法类型齐全,在传统基于卷积神经网络(convolutional neural network, CNN)、Flow-Based以及对抗生成网络(generative adversarial network, GAN)等架构基础之上,引入了最新Diffusion的架构。3)选用主流的16种评测方法对CHN-DF数据集进行综合实验,搭建面向中文场景的量化评估基准,验证CHN-DF数据集的多样性与实用性。

1 相关工作

AIGC发展带来的视频内容生成技术变革,增加了检测人脸伪造视频的紧迫性(蔺琛皓 等,2023),近些年来学术界和工业界的许多研究人员致力于创建人脸视频防伪检测数据集,开源了部分数据集以促进领域研究。本节将对人脸视频防伪检测数据集的现状进行梳理。

现有的人脸视频防伪检测数据集主要分为两类:一类数据集借助视觉层面的单模态伪造方法,通过修改或交换人类的面部特征信息达到人脸伪造的效果(丁峰 等,2024);另一类数据集伪造方法结合视觉与听觉层面的伪造手段,对于一段真实视频,通过视觉或听觉特征信息的多模态修改实现视频信息的复杂伪造,此类伪造方法伪造角度与方式多样,更贴合人脸视频恶意伪造的现实情况(李旭嵘 等,2021),但要求伪造手段多样且过程复杂,因此,此类数据集数据样本匮乏。

1.1 基于视觉的单模态人脸视频防伪检测数据集

1) UADFV (unified anti-spoofing deepfake and video dataset)(Matern等,2019)。UADFV为第1个用于人脸视频防伪检测的数据集,数据集共有98个视频,其中,49个是从YouTube收集到的真实视频,伪造视频则是通过使用FakeApp应用程序进行伪造生成49个假视频。作为早期人脸视频防伪检测数据集,UADFV在数量和质量上都有限制,由单一的FakeApp产生的假视频中,人脸扭曲变化及异常动作很明显。

2) DeepfakeTIMIT(Korshunov和Marcel,2018)。同样作为早期人脸视频防伪检测数据集,DeepfakeTIMIT的真实数据来源于32名说话人拍摄的640个视频,每个说话人视频集中包含10个高分辨率的HQ(high quality)视频和10个低分辨率的LQ(low quality)视频。假视频通过面部交换技术交换说话人间面部信息得到。然而,同样由于早期视频伪造方法的局限性,生成视频只有4 s且合成的视频往往是模糊的。

3) FF++(FaceForensics++)(Rossler等,2019)。

FF++包含来自YouTube的1 000个真实视频和4 000个基于计算机图形学和深度学习方法合成的伪造视频。数据集划分为未压缩格式和H264压缩格式，用于评估深度伪造检测方法在压缩视频和未压缩视频上的性能。然而大小和多样性不足，导致难以对由大量参数组成的高性能神经结构进行最优训练。

4）Celeb-DF（Li等，2020）。针对DeepfakeTIMIT数据质量不佳和篡改痕迹粗糙的问题，Celeb-DF（Li等，2020）对视频伪造方法进行了改进，提供了更高质量的视频。数据集中的真实视频源自YouTube中的59位说话人的590个视频，并使用改进的deepfake技术生成了5 639个虚假视频。然而，该数据集仍存在伪造方法单一的问题，不适用于现实世界中遇到的挑战。

5）DeeperForensics（Jiang等，2020）。DeeperForensics数据集中的真实视频源自100名付费演员的录制，其中采用了FF++中的视频作为面部交换伪造方法的1 000个目标视频。通过将每个源身份与10个目标视频进行面部交换，合成了1 000个假视频。此外，DeeperForensics并没有采用其他的合成方法，而是利用7种扰动方法对真实视频和伪造视频进行数据增强以增加多样性。通过这种方式创建了50 000个真实视频和10 000个伪造视频。虽然数据量明显大于早期的人脸视频防伪检测数据集，并且更具多样性，但是DeeperForensics还没有像其他数据集一样对当前人脸伪造技术广泛的评测，因此DeeperForensics的学术效能尚未完全建立。

6）WildDeepfake（Zi等，2021）。面对早期人脸视频防伪检测数据集存在缺少内容多样性和视频源低质量的问题，WildDeepfake从互联网上收集真实和深度伪造的样本，包含了视频中提取的面部动作序列，在人工去除没有对应真实人脸的视频后，真实视频数量为3 805个，伪造视频数量为3 509个。视觉效果更贴合真实生活场景，但数据量不足导致在训练高性能神经网络结构时存在局限。

7）ForgeryNet（He等，2021）。ForgeryNet是现有基于视觉的人脸视频防伪检测数据集中最大规模的数据集，提出包括时序伪造定位、空间伪造定位等多项任务，ForgeryNet采用8种深度伪造方法，生成121 617个伪造视频。视频总量达到221 247个，并且视频带有丰富的数据标注。

8）DF40（Yan等，2025）。DF40采用23种伪造方法完成了基于面部交换和面部替换技术的数据伪造，每类伪造方法的伪造数量在1 500例左右。在数据种类上DF40实现了广泛人脸视频伪造数据的构建，然而其伪造框架仍局限于CNN、GAN等传统伪造框架，在基于Diffusion的数据构建上还是空白。

1.2　基于视听的多模态人脸视频防伪检测数据集

1）DFDC（deepfake detection challenge）（Dolhansky等，2021）。DFDC是第1个在视频中包含伪造音频的数据集，起初作为Facebook发布的同名DFDC竞赛的数据集，包含5 250个视频。之后，经过数据补充，真实视频达到23 654个，伪造视频数据量达到104 500个。为了保证数据集的多样性，真实视频源取自不同的环境设置，伪造视频则由8种不同的方法生成。听觉模态上仅进行音频交换，并没有使用音频伪造方法。标签仅包含真假两个类别，没有区别伪造视频中视觉伪造与听觉伪造。

2）KoDF（Korean deepfake）（Kwon等，2021）。KoDF包含采用6种伪造方法伪造的175 776个假视频和62 166个真实视频。视频中403个说话人大多是韩国人，是为了平衡在现有的防伪数据集中亚洲人口数据不足的首次努力。然而KoDF在处理视觉与听觉信息时仅进行音频与人脸唇部动作的同步伪造，并没有使用声音克隆、声音转换等深度语音伪造方法。

3）FakeAVCeleb（Khalid等，2022）。FakeAVCeleb是首个同时包含伪造视频和伪造音频的人脸视频防伪检测数据集，从VoxCeleb2数据集选择了500个真实视频，利用了Faceswap，DeepFaceLab和FSGAN伪造面部信息，利用SV2TTS（real-time voice cloning）伪造音频，使用Wav2Lip完成音频与人脸唇部动作伪造，生成了19 500个伪造视频。

1.3　人脸视频防伪测评基准

现有人脸视频防伪数据基准总体上按照数据采集、伪造样本生成、质量评估以及人脸伪造鉴别算法评测4部分构成。

然而，人脸视频防伪领域内的数据集（表1）普遍存在基准流程不完备的问题。UADFV与DeeperForensics仅限于数据集构造，缺乏人脸伪造鉴别算法评测实验。在数据采集方面与伪造样本生成方面，UADFV、WildDeepfake、Celeb-DF以及DeeperForensics存在数据生成单一或不明确的问题，导致数据评测基准难以有效反映现实中复杂伪造场景，不

利于人脸伪造鉴别算法学习到高效鲁棒的鉴伪特征。在质量评估方面，目前仅有Celeb-DF以及DF40对所提数据进行质量的评估，然而领域内数据集进行数据筛选的工作目前依赖于人工筛查，受限于数量庞大且内容逼真的数据特性，仅依靠人工筛查机制而没有明确的数据质量评价指标进行约束难以完成高效的数据筛选。在人脸伪造鉴别算法评测方面，领域内的评测方式局限于“在当前数据集上训练，在当前数据集上测试”的方式，此类评测基准难以测评人脸伪造鉴别算法的跨域检测性能（即面对不同数据分布、不同数据质量）。此外，评测指标局限在准确率以及AUC（area under the curve）得分等单一指标上，难以满足面对数据不均衡或需要多元化显示人脸伪造鉴别算法性能时的需求。

因此，现有人脸视频防伪数据基准在数据采集、伪造样本生成、质量评估以及人脸伪造鉴别算法评测等方面存在明显的不足与缺陷，限制了高效鲁棒的人脸视频防伪算法研究。

表1　人脸视频深度防伪数据集汇总

Table 1　Summary of face video deepfake detection datasets

数据集	类型	发布年份	真实视频数量	伪造视频数量	视频总数	说话人总数	伪造方法数量	包含Diffusion	面向群体
UADFV	视频	2018	49	49	98	49	1	×	欧美
DeepfakeTIMIT	视频	2018	640	320	960	32	2	×	欧美
FF++	视频	2019	1 000	4 000	5 000	–	4	×	欧美
Celeb-DF	视频	2019	590	5 639	6 229	59	1	×	欧美
DeeperForensics	视频	2020	50 000	10 000	60 000	100	1	×	欧美
WildDeepfake	视频	2020	3 805	3 509	7 314	–	–	×	欧美
ForgeryNet	视频	2021	99 630	121 617	221 247	5 400+	8	×	欧美
DF40	视频	2024	–	31 650	31 650+	–	23	×	欧美
DFDC	视频 + 音频	2020	23 654	104 500	128 154	960	8	×	欧美
KoDF	视频 + 音频	2021	62 166	175 776	237 942	403	6	×	韩国
FakeAVCeleb	视频 + 音频	2022	500	19 500	20 000	500	4	×	欧美
CHN-DF	视频 + 音频	2025	213 187	221 540	434 727	2 540	12	√	中国

注：“–”表示该信息在原始论文与开源可获取平台中均未提供。“√”和“×”分别表示包括和不包括。

2　CHN-DF数据集

CHN-DF人脸视频防伪检测数据集包含视觉与听觉两个模态的信息。本节首先介绍CHN-DF数据集的真实视频获取和伪造视频生成，然后详细描述CHN-DF数据集的基本属性信息。图1和图2分别为CHF-DF数据流程图和CHN-DF伪造视频生成示例。

2.1　真实视频

为了保障CHN-DF数据集的场景多样性与内容复杂性，CHN-DF真实视频源于目前最大的公开中文视听多模态数据集CN-CVS（Chinese mandarin audio-visual dataset）（Chen等，2023）以及中文唇语数据集CMLR（Chinese mandarin lip reading）（Zhao等，2020）。CN-CVS总共有超过2 500名说话人，数据总条数超过20万，总时长超过300 h，CHN-DF选取其中Speech部分的2 529名说话人视频，选取的视频总量接近20万；CMLR数据集包含2009年6月至2018年6月的新闻联播视频，数据集包含由11位主持人所表述的共102 076个视频，CHN-DF数据集对CMLR数据集进行了筛选，达到保持说话人之间视频数据量平衡的目的，选取的视频总量接近2万。

基于此，CHN-DF真实视频数据量达到213 187个，超过目前公开的人脸视频防伪检测数据集的真实视频数量，说话人总数达到2 540位。此外，CMLR使用基于方向梯度直方图（histogram of oriented gradient，HOG）的人脸检测方法，再利用开源平台进行人

图 1　CHF-DF 数据流程图

Fig. 1　Data flow diagram of CHF-DF

图 2　CHN-DF 伪造视频生成示例

Fig. 2　Examples of forged videos generated in CHN-DF

脸识别和对齐；CN-CVS 使用 dlib 工具包对每个视频进行面部检测，删除没有人脸或多个人脸的视频，因此 CHN-DF 视频区域已固定在人脸部分。

由于 CHN-DF 数据集中真实视频基于说话人身份进行视频内容划分，数据集中训练集、验证集和测试集的说话人不存在重叠部分。因此，CHN-DF 数据集具有高度可扩展性。可实现将新说话人的真实视频与伪造视频加入数据集，以增加真实和深度假视频的数量，并确保训练集、验证集和测试集相互独立。

2.2　伪造视频

在视觉层面，为了能够还原日益泛滥的面部伪造场景，本文全面采用了面部交换、面部重现以及文生视频 3 类视觉层面的伪造方法，涵盖了现实社交媒体中通过“换脸”、“恶意拼接”以及提示词生成的虚假视频场景。在听觉层面，为了能够为伪造鉴别研究提供真实的语音诈骗场景，本文采用语音转换以及语音克隆两类听觉层面的伪造方法，模拟现实场景中“音色克隆”以及“通话修改”场景。

此外，随着多模态融合技术的发展，基于视听融合的多模态视频伪造方式也在互联网环境中大量传播。此类伪造视频实现了伪造语音片段与人物视觉表达的同步，即模仿人物声音进行虚假语音剪辑或合成的同时，修改视觉模态上的语言表达来适配虚假语音的片段，即“对口型”。

具体地，CHN-DF 的伪造视频构建概况如图 3 所示，由于生成的伪造视频在人工检查过程中根据伪造效果进行了筛选，因此每种伪造方法的视频数量并不相等。下面给出每类伪造方法的技术分析。

1）面部交换。采用在 HuggingFace 开源使用平台类别下载量前两名的 SimSwap（Chen 等，2020）和

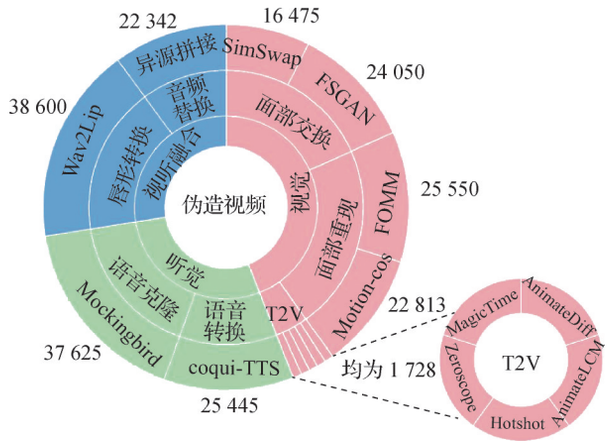

图3　CHN-DF 中不同方法伪造视频数量分布

Fig. 3　Distribution of forged video counts in CHN-DF

FSGAN（face swap generative adversarial network）（Nirkin 等，2019）算法。SimSwap 采用身份注入模块在特征级别上将源图像中人脸的身份信息转移到目标视频的人脸上，此外使用弱特征匹配损失，该损失以隐式方式帮助模型保留面部属性。FSGAN 基于对抗生成网络的换脸模型，根据目标视频和源视频能够实现面部交换。模型首先根据目标人脸的姿态和表情重新绘制源视频人脸并分割成两个面部区域，同时填补重新绘制的脸部的缺失部分并将完整的脸部与目标进行混合。

2）面部重现。采用在 Github 开源项目平台 Star 数量超过 1.5 万的 FOMM（Siarohin 等，2019）以及其续作 Motion-cos（Siarohin 等，2021）。FOMM 模型由运动估计模块和图像生成模块两个主要模块组成。根据视频中提取的帧对，将运动编码为特定于运动的关键点位移和局部仿射变换的组合，进而组合出学习运动的特征图来重建视频。Motion-cos 在此基础上实现细粒度重现编辑，基于部件分割进行自监督深度学习，从人脸源图像提取关键点信息，依据各个子部件的特征图对目标视频进行逐帧伪造，实现面部替换的区域化操作。

3）文生视频（text to video，T2V）。文生视频是新型视频内容生成的主要载体。本文选择在 HuggingFace 开源平台类别下载量前 5 的算法。Zeroscope（Cerspense，2023）、AnimateDiff（Lin 和 Yang，2024）、Hotshot（Mullan 等，2023）、AnimateLCM（Lin 和 Yang，2024）以及 MagicTime（Yuan 等，2025）。算法均基于 Diffusion 生成架构，通过扩散融合文本输入表征信息实现逐视频帧的全量生成。

4）唇形转换。Wav2Lip（Chung 和 Zisserman，2017）是唇形动作迁移开源项目中广泛使用的代码工程，Wav2Lip 可以将动态的视频进行唇形转换，实现唇形动作与输入语音匹配，即“对口型”。Wav2Lip 利用预先训练的唇语同步检测器使模型根据音频学习嘴唇动作，实现生成视频人物口型与输入语音同步。

5）语音克隆。Mockingbird（Babysor，2021）是目前已知可获取的开源中文实时语音克隆项目，通过不同讲话人音频信息合成虚假音频。Mockingbird 引入中文训练数据集用于训练语音合成系统，提取讲话人的声音提取音色向量，然后根据讲话人声音和音色向量加上合成器和声码器完成中文语音克隆。

6）语音转换。cogui-TTS 在 github 上获得超过 3.7 万 Star 并在中文社交平台上广泛使用。coqui-TTS 基于低资源零样本文本转语音模型，具有合成包括汉语在内的多种语言能力，提供了包括 Glow-TTS（Kim 等，2020）在内的多种文本语音规范模型，以及 Multiband-MelGAN（Yang 等，2021）等声码器模型。这些模型的高效性和多功能性使得 coqui-TTS 能够处理文本到语音转换任务。

7）音频替换。音频替换类别是指以异源拼接的人工方式，将源视频的音频替换为同一子集（训练集、验证集或测试集）下其他视频的音频后拼接生成的人脸伪造视频。

2.2　质量评估

为了保证 CHF-DF 的数据质量，本文采取了智能评测与人工审查结合的方式进行数据筛选。具体地，本文首先采用 YOLO（you only look once）算法对生成的伪造视频进行人脸识别，YOLO 算法识别人脸的概率低于 80% 的会被删除；接着基于 FID（Fréchet inception distance）得分，将 FID 得分高于 30 的低质量数据进行删除；最后召集具备资深计算机视觉处理经验的志愿者对数据进行人工筛选，删除存在明显视觉缺陷的视频数据。

2.3　数据集描述

CHN-DF 数据集包含 434 727 个人脸视频，说话人总数达到 2 540 人。其中，真实视频 213 187 个，伪造视频 221 540 个，CHN-DF 正负样本达到了相对平衡。根据说话人身份，按照 7∶1∶2 的比例将 CHN-DF 视频划分为训练集（1 778 位说话人的 350 679 个视频）、验证集（254 位说话人的 22 685 个视频）和测试

集(508位说话人的52 723个视频),CHN-DF视频时长分布如图4所示,持续时间在0.36 ~355.58 s,贴合现实情况中视频时长长短不一的特点,适配于多样化人脸防伪检测使用场景,平均长度为5.12 s。视频时长集中在0 ~ 20 s,其中98.75%的片段小于20 s,99.94%的片段小于50 s。

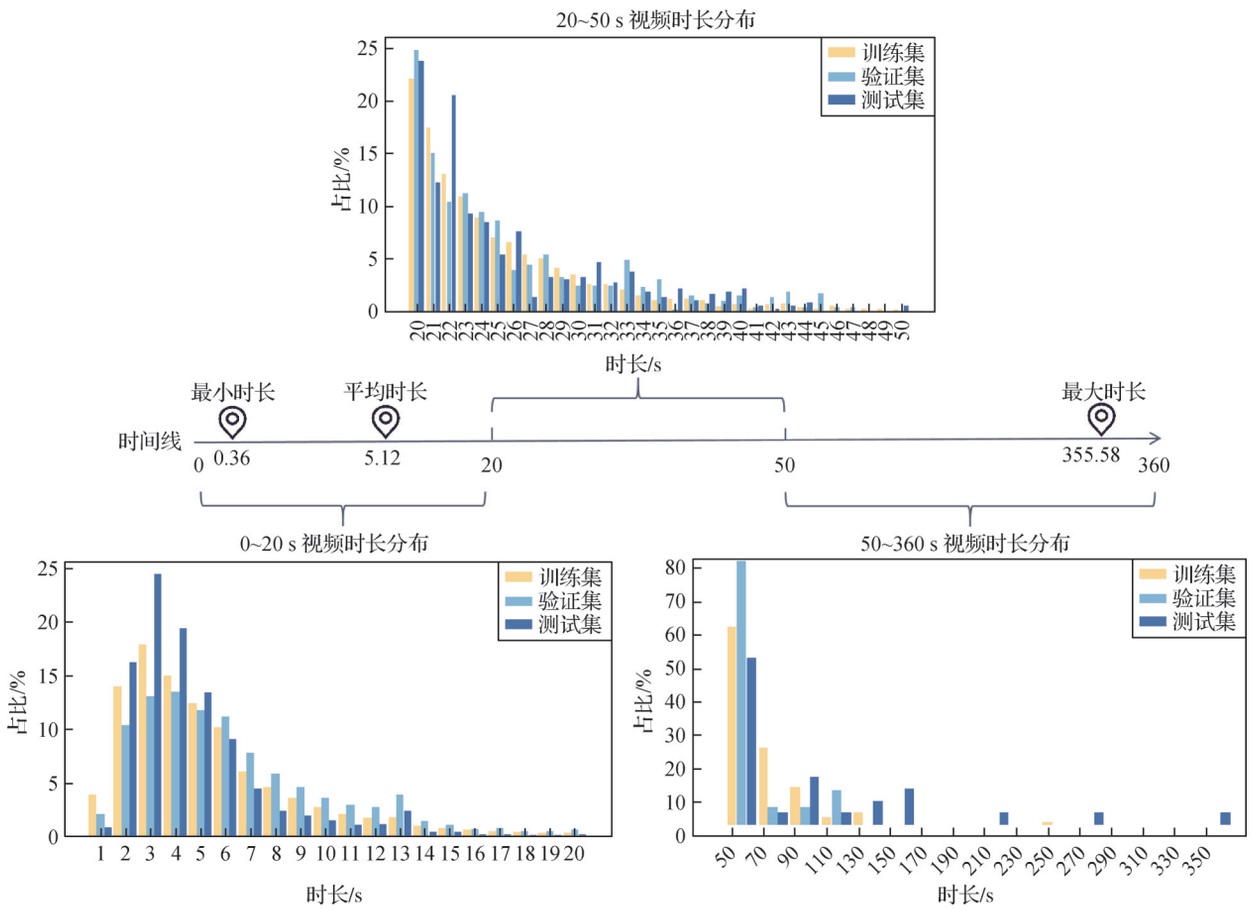

图4　CHN-DF 视频时长分布

Fig. 4　Duration distribution of videos in CHN-DF

3　CHN-DF基准评测

人脸视频防伪检测模型性能好坏通过测评模型在人脸视频防伪检测数据集的多种定量指标体现。在本节中将介绍CHN-DF基准评测的评估方法以及评价指标;基于代码的可复现性,采用16种人脸视频防伪检测领域先进方法进行全面基准性能评估,以此展示CHN-DF数据集的复杂性和贴近真实场景水平,同时与多模态FakeAVCeleb数据集进行比较。需要说明的是,选择FakeAVCeleb数据集最重要的原因是FakeAVCeleb是目前已知的唯一包含详细音视频伪造标注的多模态人脸视频防伪检测数据集。此外,该数据集还采用丰富的造假方法,在多模态人脸视频防伪检测领域是被广泛接受的优秀评测基准。

3.1　评估方法

基于数据集涵盖的视觉、听觉以及视听融合的伪造属性,CHN-DF能够为人脸视频防伪研究领域中基于单一模态防伪检测、基于集成方法的防伪检测方法以及多模态人脸视频防伪检测模型提供基准评测。具体地,在单一模态防伪检测中评测模型在面部交换、面部重现以及文生视频3类伪造方法的性能;在基于集成方法的防伪检测方法以及多模态人脸视频防伪检测模型中进行视觉、听觉以及视听融合伪造上的全面基准评测。

3.1.1　单模态方法

1)EfficientNet-B0(Tan 和 Le,2020)采用多目标神经网络架构搜索技术构建,核心模块为融合压缩激励优化的倒置瓶颈结构。通过复合缩放策略实现

网络在计算效率与检测精度间达到最优平衡。

2) CViT (convolutions to vision Transformers)(Wodajo 和 Atnafu, 2021)通过双阶段特征学习:(1)无全连接层的堆叠卷积特征提取器,专用于面部特征学习;(2)视觉Transformer编码器,将卷积特征转换为序列化像素级表征进行检测。该设计结合了CNN的局部特征提取优势与Transformer的全局建模能力。

3) F3-Net (frequency in face forgery network)(Qian等,2020)基于双流协作学习的频率感知检测框架,创新性地融合:(1)频率域分解图像成分分析;(2)局部频率统计特征建模。通过多尺度频率特征协同挖掘,有效捕获细微伪造痕迹。

4) SLADD (self-supervised learning of adversarial deepfake detecton)(Chen等,2022)通过"伪造参数配置池"技术提升检测器的场景泛化能力。其核心设计包含两大机制:(1)基于多样化伪造配置的增强样本合成系统,通过动态生成具有不同伪造特征的训练数据来提升样本多样性;(2)参数配置预测训练框架,通过要求模型同时预测输入样本的伪造配置参数,强化检测器对伪造特征的敏感性。

5) AltFreezing(Wang等,2023)针对现有模型过度依赖空间伪影的问题,提出动态权重冻结策略。将网络参数划分为空间特征相关和时序特征相关两个子集,在训练过程中通过交替冻结机制强制模型均衡学习时空特征。同时引入视频级数据增强方案,通过时序维度扰动进一步提升模型泛化性能。

6) TALL (thumbnail layout for deepfake video detection)(Xu等,2023)通过创新的视频重构方法保持时空特征完整性。具体实现包括:(1)固定位置帧掩码技术,在每帧相同位置进行特征遮蔽以增强泛化能力;(2)时序—空间特征重组模块,将处理后的视频帧重排为预设的缩略图布局,形成包含完整时空特征的检测用表征。

7) Exposing(Ba等,2024)基于信息瓶颈原理的深度伪造检测系统,其核心创新在于:(1)多粒度局部表征提取模块,捕获非重叠区域特征;(2)语义融合单元,将局部特征聚合为全局语义表征。该设计实现了从局部异常到全局伪造判定的有效映射。

8) LSDA (latent space data augmentation) (Yan等,2025)针对特定伪造痕迹过拟合问题,提出潜在空间增强策略:(1)通过潜在空间扰动扩展伪造特征空间;(2)构建平滑决策边界实现跨域泛化。该方案有效提升了不同伪造类型间的特征区分度,在跨数据集测试中表现优异。

3.1.2 集成方法

1) Meso-4(Afchar等,2018)是基于图像噪声中段信息的人脸伪造检测算法,有效解决了图像噪声减弱和高层语义特征难以区分伪造视频帧的问题。其浅层结构增强了对中等和大尺度特征的敏感度,提升了面部特征检测的能力。

2) MesoITN-4(Afchar等,2018)架构的灵感来自于InceptionNet,它通过用InceptionNet的模块替换第一层卷积层来改进Meso-4,能够更有效地捕捉不同尺度上的特征。

3) Xception(Chollet,2017)是基于深度可分离卷积层的卷积神经网络体系结构,对解耦通道相关性和空间相关性进行简化,推导出深度可分离卷积,能够高效地提取图像和视频帧中的复杂特征。

3.1.3 多模态方法

1) Multimodal-2(Verma等,2021)是一款开源的多模态模型。该模型由3部分组成:(1)处理电影海报的卷积神经网络(CNN)块,负责视觉模式;(2)处理电影类型的长短期记忆(long short-term memory, LSTM)块,负责文本模式;(3)一个前馈网络,负责分类,它综合了前两个模块的输出。在伪造视频检测中,模型利用CNN块分析视频帧的细微差异和LSTM块处理音频时序信息,有效捕捉伪造视频不一致性。

2) CDCN (central difference convolutional network)(Yu等,2020)基于中心差分卷积网络,用于解决人脸反欺骗的任务。该模型采用3层融合特征(低、中、高)预测灰度面部深度。与传统的卷积神经网络相比,CDCN通过其中心差分卷积网络能够有效地提取皮肤纹理、表情细节等细微的局部特征,有助于捕获伪造技术产生的微小瑕疵。

3) MDS (modality dissonance score) (Chugh等,2020)音画同步是伪造视频难以成功伪造的,因为被伪造的视频帧往往会存在失去唇型或不自然的面部和嘴唇运动情况。MDS比较伪造视频的视觉和听觉内容,通过量化模态之间的不协调性进行多模态伪造视频检测。

4) VFD (voice-face homogeneity tells deepfake) (Cheng等,2023)关注人的生物特征(声音和面部)

之间的匹配程度，利用了人类生物特征的内在相关性进行人脸防伪检测，学习面部和音频本质特征，拉近匹配的音视频，分离不匹配的音视频。

5）AVoiD-DF（Yang等，2023）是一种基于视听联合学习的人脸视频防伪检测方法，用于多模态人脸视频伪造检测。它由3个关键部分组成，包括时空编码器、多模态联合解码和跨模态分类器，旨在通过深度伪造在时空层次上的视听不一致性进行伪造检测。

3.2 评估指标

为了评估人脸视频防伪检测模型在数据集上的性能优劣，本文采用准确度（Accuracy）、精确度（Precision）、召回率（Recall）和F1分数（F1-score）4项指标进行性能优劣的量化客观评估，不仅考虑到4项评估指标在分类领域使用的广泛性，而且在人脸视频防伪检测这一特定领域中，这些指标的组合利于全面评估模型性能、增强防伪问题场景的关注、处理类别不平衡以及反映数据集质量。

1）全面评估模型性能。Accuracy衡量模型正确预测的总体比例，对于整体性能提供全面的视角，适用于平衡的数据集。F1-score则将精确度和召回率结合，可以在正负样本之间取得平衡，在数据集存在类别不平情况下，F1-score是一个综合的度量。

2）增强防伪问题场景的关注。在人脸视频防伪检测这类安全防护场景中，更关心的是模型对真实正例的捕获程度，即模型的结果有多少是真正例，Precision与Recall指标可以直观地反映人脸视频防伪检测模型结果在真正例结果上的优劣情况。

3）处理类别不平衡。在人脸视频防伪检测模型中，人脸视频防伪检测数据集存在类别不平衡的情况，F1-score适用于对模型性能评估时数据集类别不平衡的情况，可以更好地反映模型对正例的分类性能。

4）反映数据集质量。凭借能够衡量模型正确分类比例的属性，Accuracy对于正负样本平衡的数据集，可以作为对整体数据集质量的一个反映。在类别不平衡的情况下，F1-score可以更敏感地反映模型对少数类别的处理能力，从而更好地评估数据集的质量。

3.3 基准实验与结果分析

3.3.1 数据集预处理

为了训练基准人脸视频防伪检测模型，分别按照视觉与听觉两个模态对数据集进行预处理。对于视觉模态，由于CHN-DF的数据源CMLR和CN-CVS视频区域已固定在人脸部分，因此无需进行视频的人脸检测与定位操作，从每个视频中提取视频帧并分别存储它们，然后将视频帧作为模型输入，提取用于分辨视频真伪的视觉特征。对于听觉模态，首先按照采样率为16 kHz从视频中提取音频并以WAV格式存储。接着使用10 ms窗口位移单位的25 ms海宁（Hanning）窗口计算梅尔倒谱系数（Mel frequency cepstral coefficents，MFCC）特征。因此获得了每个音频帧包含80个MFCC特征的二维阵列（$D = 80$），将所得到的MFCC特征存储为一个三通道图像，然后对MFCC特征图像的数量进行上采样来解决每个视频只有一个MFCC特征图像的问题。将这些MFCC图像作为输入传递给模型，提取语音特征，以学习真伪语音之间的区别。

3.3.2 基准实验设置

为了CHN-DF基准评测的公平性，CHN-DF中基准人脸视频防伪检测模型采用与评测FakeAVCeleb相同的模型参数。具体地，对每种方法进行了50次迭代的训练，使用了EarlyStopping机制，其中的patience设置为10。采用了Adam优化器，学习率为10^{-5}，实验在一台搭载Silver 4310 CPU以及Nvidia A40 GPU的计算机上运行。其中在集成方法中，使用硬投票（hard-voting）和软投票（soft-voting）机制对音频和视频防伪模型进行预测结果投票集成。

3.3.3 人脸视频防伪方法对比实验

本文选择FakeAVCeleb作为对比数据集，原因如1.2节所述，在基于视觉与听觉的多模态人脸视频防伪检测数据集中FakeAVCeleb是目前已知唯一公开可获得的同时拥有深度伪造音频和深度伪造视频的数据集。因此为了进行面向人脸视频防伪检测的基准评测模型性能分析和通过实验验证CHN-DF数据集的复杂性和贴近真实场景水平，将所选方法在CHN-DF与FakeAVCeleb上进行基准评测，性能结果如表2所示。

1）CHN-DF复杂性和贴近真实场景水平验证。16种人脸视频防伪检测的基准评测模型（区分集成方法中硬投票和软投票机制）的共76项指标结果中，在CHN-DF中的指标结果共有50项低于在FakeAVCeleb中的指标结果。且基准人脸视频防伪检测模型在CHN-DF的4种指标结果视觉单模态上

表2　CHN-DF于FakeAVCeleb检测难度对比

Table 2　Comparison of detection difficulty between CHN-DF and FakeAVCeleb

防伪检测方法	年份	模态	CHN-DF				FakeAVCeleb			
			Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
EfficientNet-B0	2020	视觉	0.422 7	0.467 4	0.480 1	0.473 6	**0.433 0**	**0.504 3**	**0.490 7**	**0.497 4**
CViT	2021	视觉	0.532 2	0.578 9	**0.554 1**	**0.566 2**	**0.599 0**	**0.580 1**	0.500 0	0.537 0
F3-Net	2020	视觉	0.528 8	0.544 0	0.522 9	0.533 2	**0.570 6**	**0.582 1**	**0.566 4**	**0.574 1**
SLADD	2022	视觉	0.682 1	**0.657 7**	**0.653 2**	**0.655 4**	**0.704 5**	0.628 9	0.600 2	0.614 2
AltFreezing	2023	视觉	0.734 7	0.688 7	0.706 3	0.697 3	**0.756 7**	**0.777 8**	**0.768 9**	**0.773 3**
TALL	2023	视觉	0.754 2	**0.733 6**	**0.744 8**	**0.739 1**	**0.766 4**	0.722 4	0.733 8	0.728 0
Exposing	2024	视觉	0.804 2	0.866 3	0.817 9	0.841 4	**0.836 0**	**0.874 3**	**0.899 1**	**0.886 5**
LSDA	2024	视觉	**0.846 8**	**0.852 1**	0.863 3	**0.857 6**	0.834 2	0.844 9	**0.866 9**	0.839 5
Meso-4_S	2021	视觉 + 听觉	**0.568 5**	0.475 4	0.472 9	**0.474 1**	0.459 3	**0.537 3**	**0.510 7**	0.377 5
Meso-4_H	2021	视觉 + 听觉	**0.499 6**	0.511 9	0.509 6	**0.479 3**	0.459 3	**0.537 3**	**0.510 7**	0.377 5
MesoITN-4_S	2021	视觉 + 听觉	0.645 5	0.711 7	0.654 1	0.681 6	**0.728 7**	**0.744 5**	**0.741 9**	**0.728 6**
MesoITN-4_H	2021	视觉 + 听觉	0.581 1	0.582 3	0.533 7	0.556 9	**0.728 7**	**0.744 5**	**0.741 9**	**0.728 6**
Xception_S	2021	视觉 + 听觉	0.436 0	**0.268 6**	**0.516 3**	**0.353 3**	**0.439 4**	0.219 7	0.500 0	0.305 2
Xception_H	2021	视觉 + 听觉	0.436 0	**0.268 6**	**0.516 3**	**0.353 3**	**0.439 4**	0.219 7	0.500 0	0.305 2
Multimodal-2	2021	视觉 + 听觉	0.502 0	0.251 0	0.500 0	0.334 2	**0.674 0**	**0.679 0**	**0.673 5**	**0.671 5**
CDCN	2021	视觉 + 听觉	0.500 0	0.500 0	0.500 0	**0.467 8**	**0.515 0**	**0.500 0**	**0.500 4**	0.385 5
MDS	2020	视觉 + 听觉	0.578 4	**0.857 1**	0.452 1	0.591 9	**0.690 0**	0.780 0	**0.695 0**	**0.665 0**
VFD	2022	视觉 + 听觉	0.643 9	0.711 3	0.654 4	0.681 6	**0.815 2**	**0.837 7**	**0.754 2**	**0.793 7**
AVoiD-DF	2023	视觉 + 听觉	0.695 7	0.724 4	0.678 5	0.700 6	**0.837 1**	**0.841 1**	**0.770 2**	**0.804 0**

注：加粗字体表示同一防伪检测方法在CHN-DF和FakeAVCeleb各指标上的最佳指标性能；集成方法中“_S”与“_H”分别表示软投票和硬投票机制。

低于0.87，在视听融合的模态集中为0.6以下。在侧重关注防伪问题场景的Precision与Recall指标，以及分别适用于正负样本平衡与失衡的Accuracy和F1-score上，人脸视频防伪检测基准评测模型在CHN-DF中性能相较于在FakeAVCeleb上表现不佳，面临的防伪任务更复杂、更具挑战性。由此也验证了CHN-DF数据集的复杂性和更贴近真实场景水平，更有利于推动性能更好的深度防伪检测方法的研发。

2）面向人脸视频防伪检测的评测模型性能分析。AVoiD-DF与LSDA在CHN-DF和FakeAVCeleb中，分别在视觉单模态与视听模态兼备的数据中取得了最佳性能结果，防伪效果最优。可能的原因是AVoiD-DF相较于其他基准人脸视频防伪检测模型，在基于视听联合学习模块引入的多模态联合解码MMD（multimodal joint decoding）中，使用MMD模块进行模态融合。相较于其他多模态方法模态融合结构，AVoiD-DF中输入的视觉和音频嵌入块是通过两个并行解码器通道馈送，每个通道都有一个双向交叉注意（BiCroAtt）模块，且之后有自注意力块和前馈层搭配。这使得两种模态之间具备更好的信息共享与联合学习能力。对于LSDA，模型通过扩展伪造特征空间，在潜在空间内构建并模拟伪造特征的变化，从而学习更具泛化性的决策特征。该过程有助于提取丰富的领域特异性特征，并促进不同伪造类型间更平滑的过渡，有效弥合域间差异。然而AVoiD-DF在CHN-DF的指标结果明显低于在FakeAVCeleb上的结果，可能的原因是AVoiD-DF作为基于视听联合学习的人脸视频防伪检测方法，在面对伪造视觉—伪造听觉（V_FA_F）情况时（如Wav2Lip将动态的视频

进行唇形转换，实现唇形动作与输入语音匹配的视频）面部与音频的内在相关性会被破坏，同时 CHN-DF 相较于 FakeAVCeleb 在 V_FA_F 中采用更为复杂的伪造手段，因此 AVoiD-DF 在视听伪造信息更为复杂的 CHN-DF 中面对 V_FA_F 类别数据时指标结果较低。

MesoITN-4 在基于集成方法的人脸视频防伪检测基准评测模型中防伪效果最优，可能的原因是 MesoITN-4 针对伪造视频中伪造方法只能合成有限分辨率的人脸图像并且必须对其进行仿射变换以匹配源人脸的配置这一视频属性，使用变体 inception 模块关注仿射变换中扭曲面区域和周围环境的分辨率不一致而产生的伪影。然而 MesoITN-4 在处理采用 FOMM 和 Motion-cos 等基于面部重现的伪造视频时，由于面部重现技术并不仅是将人脸区域进行仿射变换，面部重现更注重通过保留目标人物的身份来应用源人物的特征，而面部交换更注重在两个图像之间进行面部特征的交换。因此面部重现技术产生的伪影并不等同于面部交换过程中产生的伪造痕迹，MesoITN-4 在处理通过 FOMM 和 Motion-cos 生成的伪造视频存在算法设计层面的局限性，导致基准评测的指标结果较低。

Multimodal-2 与 Xception 在 CHN-DF 和 FakeAVCeleb 中指标结果较低，在 CHN-DF 中各项指标结果在 0.52 以下，造成这种结果的一个可能原因是 Multimodal-2 与 Xception 是计算机视觉领域通用分类模型，在各种分类任务中能够取得良好的结果，但可能是由于其预训练权重和特定任务之间的领域差异，而不一定适用于视频数据中的复杂特征和动态变化。另一方面，由于人脸视频防伪检测任务涉及到更丰富的信息，包括面部表情、姿势等因素，这可能导致了通用分类模型在该任务上的性能不佳。

此外，在面向人脸视频防伪检测的基准评测模型中多模态方法优于集成方法的性能结果，可能的原因是相较于集成方法中多个单模态分类器模型组成整体模型的思路，多模态方法在处理人脸视频伪造数据时考虑到视觉与听觉之间的相关性与一致性信息。相对于单模态（视觉或听觉）的伪造，伪造方法在篡改视觉与听觉之间相关性的特征时难度更大，使得伪造的效果更易于捕捉，所以视觉与听觉之间的相关性特征能够为人脸视频防伪检测模型提供更明显的检测特征，因此处理具备多模态信息的人脸视频伪造中多模态方法防伪检测效果更优。

3.3.4 跨数据集防伪方法对比实验

为了评估 CHN-DF 数据集的质量和衡量基准人脸视频防伪检测模型的泛化性能，进行跨数据集防伪方法对比实验。实验使用基准模型在 FakeAVCeleb 上进行训练并在 CHN-DF 上进行测试。通过在 FakeAVCeleb 上进行训练，模型能够学习人脸伪造视频的数据分布，在 CHN-DF 上进行测试能够提供模型在与训练集不同分布数据中的性能表现，有助于验证模型在面对未知数据时的鲁棒性和泛化性，同时在 FakeAVCeleb 上的训练模型与在 CHN-DF 上的训练模型的测试结果对比也可评估 CHN-DF 数据集的质量。

16 种人脸视频防伪检测评估模型在以 FakeAVCeleb 为训练集并以 CHN-DF 为测试集的跨数据集防伪任务中各项指标明显降低，表明模型在 CHN-DF 中面对更复杂和更具挑战性的伪造数据，由此进一步验证了 CHN-DF 数据集的复杂性和贴近真实场景水平。表 3 展示了跨数据集防伪方法对比实验结果。

结合表 2 在 CHN-DF 数据集多模态防伪方法对比实验结果，可以发现由于数据集之间数据的来源不同，在跨数据集的防伪任务中，16 种人脸视频防伪检测模型性能指标均有明显的下降。其中，MesoITN-4 指标结果下降最为显著，可能的原因是 MesoITN-4 在 FakeAVCeleb 中缺少基于面部重现的伪造视频的训练，导致通过捕捉伪影进行视频防伪检测的局限更加明显；VFD 在跨数据集的防伪任务中指标虽有下降但取得最优的防伪效果，可能的原因是 VFD 的微调（fine-tune）机制基于预训练模型进行微调，因此快速适应新的任务或数据集；Multimodal-2、Xception 以及 MDS 在跨数据集的防伪任务中指标下降幅度较低，可能的原因是 Multimodal-2 与 Xception 作为通用分类模型虽然不一定适用于视频数据，但 Multimodal-2 与 Xception 良好泛化性能使得模型在跨数据集任务中指标波动幅度降低。

3.3.5 中文场景人脸伪造鉴别方法性能评测

为了对 CHN-DF 的独特价值以及创新进行实验性说明，以训练数据是中文场景（CHN-DF）为实验组，训练数据不是中文场景（CHN-DF）为对照组，探究该影响因素下模型面对中文伪造场景时识别性能的差异。具体地，选择 FakeAVCeleb（英语体系）、CHN-DF（中文体系）、DFDC（英语体系）以及 KoDF

表3　以FakeAVCeleb为训练集并以CHN-DF为测试集的跨域评测结果

Table 3　Cross-Domain evaluation results using FakeAVCeleb as the training set and CHN-DF as the test set

防伪检测方法	模态	Accuracy	Precision	Recall	F1-score
EfficientNet-B0	视觉	0.416 5	0.475 2	0.450 0	0.462 2
CViT	视觉	0.514 1	0.520 2	0.500 0	0.509 9
F3-Net	视觉	0.503 5	0.534 3	0.544 8	0.539 4
SLADD	视觉	0.598 7	0.564 3	0.602 1	0.582 5
AltFreezing	视觉	0.654 1	0.637 4	0.603 2	0.619 8
TALL	视觉	0.689 8	0.700 1	0.695 8	0.697 9
Exposing	视觉	**0.715 2**	0.723 6	0.704 4	0.713 8
LSDA	视觉	0.706 5	**0.733 6**	**0.711 7**	**0.722 4**
Meso-4_S	视觉 + 听觉	0.400 7	0.384 4	0.499 8	0.434 5
Meso-4_H	视觉 + 听觉	0.413 5	0.332 1	0.446 3	0.380 8
MesoITN-4_S	视觉 + 听觉	0.411 7	0.410 0	0.413 3	0.411 6
MesoITN-4_H	视觉 + 听觉	0.400 2	0.391 1	0.403 5	0.397 2
Xception_S	视觉 + 听觉	0.397 1	0.213 4	0.429 9	0.285 2
Xception_H	视觉 + 听觉	0.397 1	0.213 4	0.429 9	0.285 2
Multimodal-2	视觉 + 听觉	0.414 5	0.342 3	0.399 7	0.368 7
CDCN	视觉 + 听觉	0.378 4	0.331 2	0.452 1	0.382 3
MDS	视觉 + 听觉	0.522 3	**0.648 7**	0.403 3	0.497 3
VFD	视觉 + 听觉	**0.601 1**	0.587 7	**0.530 1**	**0.557 4**
AVoiD-DF	视觉 + 听觉	0.599 7	0.600 3	0.498 3	0.544 5

注：加粗字体表示在视觉模态和视觉+听觉模态上的最佳指标性能。

（韩文体系）作为训练数据，探究在不同国别体系数据下训练的模型在真实中文互联网人脸视频伪造场景中的性能表现。

值得注意的是，为了保证评测的公平性：1）选择目前领域内所有多模态人脸视频防伪数据集参与对比评测：FakeAVCeleb、CHN-DF、DFDC以及KoDF。其中，数据集数据的选择原则为伪造方法的交集部分，即采用相同人脸伪造方法生成的视频作为训练数据，达到控制中文数据为唯一变量的目的。2）测试场景中的中文数据覆盖超过1 000个场景与人物，且人物和场景上与CHN-DF的训练数据不存在重叠情况（如2.1节所述）。因此在探究CHN-DF对于中文场景的人脸伪造鉴别方法模型的贡献中不存在数据不公平性。

实验结果如表4所示，在中文伪造的人脸视频检测中，CHN-DF的使用使得19种仿伪检测方法中17个达到了最优，2个达到了次优效果。KODF（韩语）数据集因其韩国群体的特性，在面部视觉上与中国群体相近，因此在视觉模态的检测中效果总体上次优。然而在听觉模态，由于汉语、英语以及韩语语义信息表达上的差异，其他语种的训练数据无法对中文场景的人脸视频检测提供有效帮助，由此体现出构建面向人脸视频防伪检测的大规模中文数据测评基准CHN-DF，对中文场景的人脸视频防伪研究的独特价值与重要贡献。

4　面临挑战与发展方向

本文构建的首个面向人脸视频防伪检测的大规模中文数据评测基准，在真实性、多样性、准确性和对抗性等方面仍存在诸多挑战，针对这些挑战，构建更优质的数据评测基准数据集，对于推动深度防伪检测技术高质量发展有着重要的意义。基准评测数据集构建当前存在的主要局限如下：

表 4　中文场景训练时中文伪造视频识别准确性的影响实验

Table 4　Impact of training on Chinese scenarios on detection accuracy for Chinese forged videos

防伪检测方法	模态	训练集			
		FakeAVCeleb (英语)	CHN-DF (汉语)	DFDC (英语)	KoDF (韩语)
EfficientNet-B0	视觉	0.416 5	**0.422 7**	0.398 6	0.420 1
CViT	视觉	0.514 1	**0.532 2**	0.508 9	0.526 9
F3-Net	视觉	0.503 5	0.528 8	0.510 0	**0.530 1**
SLADD	视觉	0.598 7	**0.682 1**	0.6229	0.674 2
AltFreezing	视觉	0.654 1	**0.734 7**	0.7004	0.663 0
TALL	视觉	0.689 8	**0.754 2**	0.733 6	0.734 6
Exposing	视觉	0.715 2	**0.804 2**	0.725 4	0.724 9
LSDA	视觉	0.706 5	0.846 8	0.700 1	**0.850 1**
Meso-4_S	视觉 + 听觉	0.400 7	**0.568 5**	0.408 9	0.392 6
Meso-4_H	视觉 + 听觉	0.413 5	**0.499 6**	0.423 9	0.413 6
MesoITN-4_S	视觉 + 听觉	0.411 7	**0.645 5**	0.436 5	0.425 9
MesoITN-4_H	视觉 + 听觉	0.400 2	**0.581 1**	0.571 9	0.536 3
Xception_S	视觉 + 听觉	0.397 1	**0.436 0**	0.346 9	0.411 8
Xception_H	视觉 + 听觉	0.397 1	**0.436 0**	0.426 3	0.423 6
Multimodal-2	视觉 + 听觉	0.414 5	**0.502 0**	0.423 0	0.488 7
CDCN	视觉 + 听觉	0.378 4	**0.500 0**	0.356 9	0.444 4
MDS	视觉 + 听觉	0.522 3	**0.578 4**	0.427 7	0.436 9
VFD	视觉 + 听觉	0.601 1	**0.643 9**	0.573 1	0.588 9
AVoiD-DF	视觉 + 听觉	0.599 7	**0.695 7**	0.536 9	0.645 7

注：加粗字体表示每行最优结果。

1）深度伪造技术局限。AIGC 的发展使得图像生成（AI 绘画）和视频生成效果更逼真，然而现有的 AICG 伪造技术在生成长视频方面仍存在少量问题：（1）讲话人说话期间存在面部短暂性闪烁现象；（2）存在伪造面部区域边缘模糊的情况；（3）面部纹理过度平滑或缺乏细节；（4）头部姿势移动或动作不自然；（5）缺乏面部遮挡物，如眼镜、光照效果等；（6）对身体姿态或皮肤颜色一致性变化敏感，易造成身份泄露；（7）伪造视频缺乏自然的情绪和语气停顿，会出现呼吸急促、语气僵硬的现象。这种因伪造技术带来的瑕疵也同样是伪造检测技术需要关注学习的特征，但过度关注这些特征会导致伪造检测模型过拟合，在真实应用场景中鲁棒性不足。同样地，这些伪造视频的瑕疵，一定程度干扰了评测基准的准确性与客观性，但当前阶段很难避免，现有的生成技术生成结果很难保证整体自然度、流畅性与连续一致性等贴近真实场景特性。为降低生成技术不足导致的评测基准不客观，本文在构建评测基准时通过人工筛选过于明显的瑕疵数据，减少低质量对伪造检测技术定量评估成效的干扰。

2）语音数据缺乏多样性。现阶段人脸视频防伪检测领域评测基准中缺乏语音数据，在语音数据多样性方面难以保证，语音伪造技术缺乏包含多种情感表达的语音数据，使得评测基准无法充分覆盖对情感检测的测试。在多样化文化背景的语音数据收集上也面临巨大挑战，尤其是在中文数据方面，中文作为世界上使用人数最多的语言，涵盖多个方言和口音，而且不同地区和社会群体的语音表达方式各异。不同语音风格和口音数据的缺乏可能导致评测基准在应对特定口音或语音风格时的不足。因此，

构建覆盖多样性、个性化的语音数据样本，也是本文未来工作的主要方向之一。

3）标签缺乏准确性。有效的人脸视频防伪检测评测基准建立在能够贴近现实生活场景的数据集基础之上，贴近现实生活场景的数据集需要准确的标签，然而大规模的标注可能导致标签缺乏准确性，特别是在深度伪造的场景下，标注人员在标注视频数据时可能导致标签的主观性和不一致性，例如，对于AIGC创造的高质量伪造视频，人工标注会出现耗费大量时间但标签缺乏准确性的情况；在标注细粒度标签时，细粒度的标签需要标注人员对伪造技术有深入研究和专业知识，标注人员可能无法准确地识别所有细节。这些标签的主观性和不一致性情况会导致数据集在制作的过程中面临标签缺乏准确性的挑战。

4）难以抵挡对抗性攻击。人脸视频防伪检测评测基准中缺乏对抗性攻击，现实场景中攻击者在制作伪造人脸视频的同时也会考虑如何加入对抗性攻击，达到降低检测效果的目的，如刻意调整光线强度增加模型提取视觉特征的难度等，导致训练出的防伪模型容易受到对抗性攻击的影响，这些复杂场景情况在人脸视频防伪检测评测基准的构建过程中难以有效考虑并覆盖到，导致在面对现实场景时伪造检测算法面临巨大的挑战。

5 结 论

在AIGC时代人脸视频生成领域的信息存在真实性验证困难的环境下，本文提出面向人脸视频防伪检测的大规模中文数据评测基准，发布了首个面向人脸视频防伪检测的大规模中文数据集——CHN-DF，填补人脸视频防伪检测数据集大规模中文数据的空白，本文详细介绍了构建CHN-DF数据集以及中文数据评测基准的流程，并针对主流防伪检测方法进行了对比实验，从基准评测模型性能、跨数据集泛化性能等方面分析了现有人脸视频防伪检测方法的优劣。此外，大量的实验也验证了CHN-DF数据集的复杂性和贴近真实场景水平，希望面向人脸视频防伪检测的大规模中文数据评测基准，能够帮助研究人员构建性能更为优异的人脸视频防伪检测模型，成为未来人脸视频防伪检测领域研究的基石。同时，本文还指出了中文人脸视频防伪检测数据集与评测基准当前面临的挑战以及未来发展方向，希望为推动人脸视频防伪检测领域技术发展提供新的视角与方向。

版权及数据可用性声明：鉴于人脸视频数据的敏感特性，在此声明本文真实数据源（CN-CVS与CMLR）的获取途径为通过申请表单向原作者询问形式，对于数据的整合与加工（包括数据预处理以及伪造视频的合成）已获得作者同意，并无潜在版权问题。另有，CHN-DF数据集已签署CC-BY-NC协议，将数据集使用范围限定在科研教育范围内，禁止商业化使用，具体可见开源网站平台。

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 2018 IEEE International Workshop on Information Forensics and Security. Hong Kong, China: IEEE: 1-7 [DOI: 10.1109/WIFS.2018.8630761]

Babysor. 2021. Mockingbird [EB/OL]. [2025-02-10]. https://github.com/babysor/MockingBird

Ba Z J, Liu Q Y, Liu Z G, Wu S, Lin F, Lu L, et al. 2024. Exposing the deception: uncovering more forgery clues for deepfake detection [EB/OL]. [2025-02-10]. https://arxiv.org/pdf/2403.01786.pdf

Cerspense. 2023. Zeroscope [EB/OL]. [2025-02-10]. https://huggingface.co/cerspense/zeroscope_v2_30x448x256

Chen C, Wang D and Zheng T F. 2023. CN-CVS: a mandarin audio-visual dataset for large vocabulary continuous visual to speech synthesis//Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island, Greece: IEEE: 1-5 [DOI: 10.1109/ICASSP49357.2023.10095796]

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. New Orleans, USA: IEEE: 18689-18698 [DOI: 10.1109/CVPR52688.2022.01815]

Chen R W, Chen X H, Ni B B and Ge Y H. 2020. SimSwap: an efficient framework for high fidelity face swapping//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 2003-2011 [DOI: 10.1145/3394171.3413630]

Cheng H, Guo Y Y, Wang T Y, Li Q, Chang X J and Nie L Q. 2023. Voice-face homogeneity tells deepfake. ACM Transactions on Multimedia Computing, Communications, and Applications, 20(3): #76 [DOI: 10.1145/3625231]

Chollet F. 2017. Xception: deep learning with depthwise separable convolutions//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1800-1807 [DOI: 10.1109/CVPR.2017.195]

Chugh K, Gupta P, Dhall A and Subramanian R. 2020. Not made for each other- audio-visual dissonance-based deepfake detection and localization//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 439-447 [DOI: 10.1145/3394171.3413700]

Chung J S and Zisserman A. 2017. Out of time: automated lip sync in the wild//Proceedings of 2016 ACCV International Workshops on Computer Vision——ACCV 2016 Workshops. Taipei, China: Springer: 251-263 [DOI: 10.1007/978-3-319-54427-4_19]

Ding F, Kuang R S, Zhou Y, Sun L, Zhu X G and Zhu G P. 2024. A survey of Deepfake and related digital forensics. Journal of Image and Graphics, 29(2): 295-317 (丁峰, 匡仁盛, 周越, 孙珑, 朱小刚, 朱国普. 2024. 深度伪造及其取证技术综述. 中国图象图形学报, 29(2): 295-317) [DOI: 10.11834/jig.230088]

Dolhansky B, Bitton J, Pflaum B, Lu J K, Howes R, Wang M L, et al. 2020. The DeepFake detection challenge (DFDC) dataset [EB/OL]. [2025-02-10]. https://arxiv.org/pdf /2006.07397. pdf

Han Y H, Li S Y, Liu Y X, Yan Z L, Dai Y T, Yu P S, et al. 2023. A comprehensive survey of AI-generated content (AIGC): a history of generative AI from GAN to ChatGPT [EB/OL]. [2025-02-10]. https://arxiv.org/pdf/2303.04226.pdf

He Y A, Gan B, Chen S Y, Zhou Y C, Yin G J, Song L C, et al. 2021. ForgeryNet: a versatile benchmark for comprehensive forgery analysis//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 4358-4367 [DOI: 10.1109/CVPR46437.2021.00434]

Jiang L M, Li R, Wu W, Qian C and Loy C C. 2020. DeeperForensics-1.0: a large-scale dataset for real-world face forgery detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 2886-2895 [DOI: 10.1109/CVPR42600.2020.00296]

Khalid H, Tariq S, Kim M and Woo S. 2022. FakeAVCeleb: a novel audio-video multimodal Deepfake dataset [EB/OL]. [2025-02-10]. https://arxiv.org/pdf/2108.05080.pdf

Kim J, Kim S, Kong J and Yoon S. 2020. Glow-TTS: a generative flow for text-to-speech via monotonic alignment search//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 8067-8077

Korshunov P and Marcel S. 2018. DeepFakes: a new threat to face recognition? Assessment and detection [EB/OL]. [2025-02-10]. https://doi.org/10.48550/arXiv.1812.08685

Kwon P, You J, Nam G, Park S and Chae G. 2021. KoDF: a large-scale Korean DeepFake detection dataset//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 10724-10733 [DOI: 10.1109/ICCV48922.2021.01057]

Li X R, Ji S L, Wu C M, Liu Z G, Deng S G, Cheng P, e tal. 2021. Survey on deepfakes and detection techniques. Journal of Software, 32(2): 496-518 (李旭嵘, 纪守领, 吴春明, 刘振广, 邓水光, 程鹏, 等. 2021. 深度伪造与检测技术综述. 软件学报, 32(2): 496-518) [DOI: 10.13328/j.cnki.jos.006140]

Li Y Z, Yang X, Sun P, Qi H G and Lyu S W. 2020. Celeb-DF: a large-scale challenging dataset for Deepfake forensics [EB/OL]. [2025-02-10]. https://arxiv.org/pdf/1909.12962.pdf

Lin C H, Shen C, Deng J Y, Hu P B, Wang Q, Ma S Q, et al. 2023. Digitally forged face content creation and detection. Journal of Computer Science, 46(3): 469-498 (蔺琛皓, 沈超, 邓静怡, 胡鹏斌, 王骞, 马仕清, 等. 2023. 虚假数字人脸内容生成与检测技术. 计算机学报, 46(3): 469-498) [DOI: 10.11897/SP.J.1016.2023.00469]

Lin S C and Yang X. 2024. AnimateDiff-lightning: cross-model diffusion distillation [EB/OL]. [2025-02-10]. https://arxiv.org/pdf/2403.12706.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. Waikoloa, USA: IEEE: 83-92 [DOI: 10.1109/WACVW.2019.00020.]

Mullan J, Crawbuck D and Sastry A. 2023. Hotshot-XL [EB/OL]. [2025-02-10]. https://github.com/hotshotco/hotshot-xl

Nirkin Y, Keller Y and Hassner T. 2019. FSGAN: subject agnostic face swapping and reenactment//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 7183-7192 [DOI: 10.1109/ICCV.2019.00728]

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]

Rossler 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]

Siarohin A, Lathuilière S, Tulyakov S, Ricci E and Sebe N. 2019. First order motion model for image animation//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 7137-7147

Siarohin A, Subhankar R, Stephane L, Sergey T, Elisa R and Nicu S. 2021. Motion-supervised co-part segmentation//Proceedings of the 25th International Conference on Pattern Recognition. Milan, Italy: IEEE: 9650-9657 [DOI: 10.1109/ICPR48806.2021.9412520]

Tan M X and Le Q V. 2020. EfficientNet: rethinking model scaling for convolutional neural networks [EB/OL]. [2025-02-10]. https://arxiv.org/pdf/1905.11946.pdf

Verma D. 2021. Multimodal-2 [EB/OL]. [2025-02-10]. https://github.com/dh1105/Multi-modal-movie-genre-prediction

Wang F Y, Huang Z Y, Shi X Y, Bian W K, Song G L, Liu Y, et al. 2024. AnimateLCM: accelerating the animation of personalized diffusion models and adapters with decoupled consistency learning [EB/OL]. [2025-02-10]. https://arxiv.org/pdf/2402.00769v1.pdf

Wang Z D, Bao J M, Zhou W G, Wang W L and Li H Q. 2023. AltFreezing for more general video face forgery detection//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 4129-4138 [DOI: 10.1109/CVPR52729.2023.00402]

Wodajo D and Atnafu S. 2021. Deepfake video detection using convolutional vision transformer [EB/OL]. [2025-02-10]. https://arxiv.org/pdf/2103.01325.pdf

Xu Y T, Liang J, Jia G Y, Yang Z M, Zhang Y H and He R. 2023. TALL: thumbnail layout for deepfake video detection//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 22658-22668 [DOI: 10.1109/ICCV51070.2023.02071]

Yan Z Y, Yao T P, Chen S, Zhao Y D, Fu X H, Zhu J W, et al. 2025. DF40: toward next-generation deepfake detection//Proceedings of the 38th International Conference on Neural Information Processing Systems. New York, USA: ACM: 29387 - 29434 [DOI: 10.5555/3737916.3738841]

Yang G, Yang S, Liu K, Fang P, Chen W and Xie L. 2021. Multi-Band Melgan: faster waveform generation for high-quality text-to-speech//Proceedings of 2021 IEEE Spoken Language Technology Workshop. Shenzhen, China: IEEE: 492-498 [DOI: 10.1109/SLT48900.2021.9383551]

Yang W Y, Zhou X Y, Chen Z K, Guo B F, Ba Z J, Xia Z H, et al. 2023. AVoiD-DF: audio-visual joint learning for detecting Deepfake. IEEE Transactions on Information Forensics and Security, 18: 2015-2029 [DOI: 10.1109/TIFS.2023.3263748]

Yu Z T, Zhao C X, Wang Z Z, Qin Y X, Su Z, Li X B, et al. 2020. Searching central difference convolutional networks for face anti-spoofing//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5294-5304 [DOI: 10.1109/CVPR42600.2020.00534]

Yuan S H, Huang J F, Shi Y J, Xu Y Q, Zhu R J, Lin B, et al. 2025. MagicTime: time-lapse video generation models as metamorphic simulators. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47 (9): 7340-7351 [DOI: 10.1109/TPAMI.2025.3558507]

Zhao W X, Zhou K, Li J Y, Tang T Y, Wang X L, Hou Y P, et al. 2023. A survey of large language models [EB/OL]. [2025-02-10]. https://arxiv.org/pdf/2303.18223.pdf

Zhao Y, Xu R and Song M L. 2020. A cascade sequence-to-sequence model for Chinese mandarin lip reading//Proceedings of the 1st ACM International Conference on Multimedia in Asia. New York, USA: ACM: 1 - 6 [DOI: 10.1145/3338533.3366579]

Zi B J, Chang M H, Chen J J, Ma X J and Jiang Y G. 2023. WildDeepfake: a challenging real-world dataset for deepfake detection [EB/OL]. [2025-02-10]. https://arxiv.org/pdf/2101.01456.pdf

作者简介

贝毅君，男，副研究员，博士生导师，主要研究方向为大数据和工业互联网。E-mail：beiyj@zju.edu.cn

胡秉德，通信作者，男，博士，主要研究方向为数据挖掘和图神经网络。E-mail：tonyhu@zju.edu.cn

娄恒瑞，男，博士研究生，主要研究方向为深度学习和防伪检测。E-mail：hengrui@zju.edu.cn

高克威，男，博士研究生，主要研究方向为深度学习和迁移学习。E-mail：gaokw@zju.edu.cn

宋杰，男，副教授，博士生导师，主要研究方向为计算机视觉和人工智能。E-mail：sjie@zju.edu.cn

王蕊，女，研究员，博士生导师，主要研究方向为人工智能安全和计算机视觉。E-mail：wangrui@iie.ac.cn

金苍宏，男，副教授，硕士生导师，主要研究方向为大数据和智能决策。E-mail：jinch@hzcu.edu.cn

雷杰，男，讲师，硕士生导师，主要研究方向为模型优化。E-mail：jasonlei@zjut.edu.cn

宋明黎，男，教授，博士生导师，主要研究方向为计算机视觉和机器学习。E-mail：songml@zju.edu.cn

冯尊磊，男，副教授，博士生导师，主要研究方向为计算机视觉、医学影像。E-mail：zunleifeng@zju.edu.cn