Manipulated Regions Localization For Partially Deepfake Audio: A Survey

Jiayi He, Jiangyan Yi *Member, IEEE,*, Jianhua Tao* *Senior Member, IEEE*, Siding Zeng, and Hao Gu

***Abstract*—With the development of audio deepfake techniques, attacks with partially deepfake audio are beginning to rise. Compared to fully deepfake, it is much harder to be identified by the detector due to the partially cryptic manipulation, resulting in higher security risks. Although some studies have been launched, there is no comprehensive review to systematically introduce the current situations and development trends for addressing this issue. Thus, in this survey, we are the first to outline a systematic introduction for partially deepfake audio manipulated region localization tasks, including the fundamentals, branches of existing methods, current limitations and potential trends, providing a revealing insight into this scope.**

***Index Terms*—Partially deepfake audio, manipulated region localization, deepfake detection, anti-spoofing.**

I. Introduction

WITH the rapid development of artificial intelligence generated content (AIGC) techniques, significant improvements have been made in the naturalness, realism, and diversity of the synthetic audio. However, at the meanwhile, the misuse of the advanced technology may also poses a serious threat to social security, cyber security, and privacy security. In order to defend against these issues, deepfake audio detection had raised the attention in the past few years. To date, many effective countermeasures (CMs) have emerged[1–10], and the performance of some models evaluated through equal error rate (EER) are reported to be less than 1%[1, 10], indicating significant success in defending against fully deepfake audio attacks. However, with each step forward, new challenges are emerged. Partially deepfake audio attacks, a more covert way of spoofing, have attracted another round of attention. The partially deepfake audio usually combines real and fake audio clips or another real clips from other corpus, increasing the complexity and difficulty of recognizing the attacks. Existing research shows that both humans and machines can be easily deceived by partially deepfake audio[11]. To cope with this new challenge, in recent years, some fundamental facilities, such as diverse datasets, and competitions are launched to attract the attention in the community[12–17] and some CMs are proposed to locate the partially manipulations[14, 18–35]. Up to now, for the commonly used datasets, the best performance reported was a frame-level EER of 3.58%[35] on PartialSpoof datasets[12] and a segment-level F1-score of 0.7397[33] on ADD2023Track2 datasets[16]. In PartialSpoof datasets, all manipulated regions are generated by TTS/VC. However, with the development of spoofing techniques, genuine clips are also used for tampering. The ADD2023Track2 datasets takes the situation of 'truth for truth' into consideration. Besides, additional noise, format conversion and the smoothing processing on spliced traces were done. Moreover, with the widely application of large language model (LLM), advanced techniques for local feature matching and seamless stitching are bound to arrive, which will pose new challenges for partially deepfake audio localization tasks. Therefore, in order to better defend against it in the near future, we urgently need a comprehensive review to help understand the current situations, including existing outperforming CMs and the development trends of this issue.

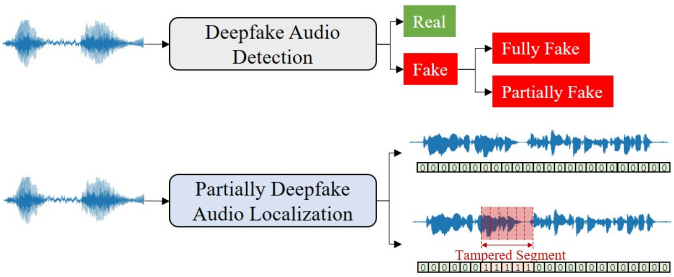


Fig. 1. The differences between detection tasks and localization tasks.

Thus, in this survey, we aim to provide a comprehensive overview of the current state in this scope, summarizing and comparing existing methodologies, and highlighting their respective strengths and weaknesses. Also, it is organized to guide future research directions and foster technological advancements by identifying gaps and challenges in the current research that remain. Ultimately, this survey aims to enhance researchers' understanding and raise community awareness of manipulated regions localization tasks for partially deepfake audio. Additionally, we also hope that it can become a guide for beginners in this research scope. The contributions of this survey are presented as following:

- This is the first comprehensive survey focusing on partially deepfake audio manipulated regions localization

Jiayi He is with the State Key Laboratory of Multi-modal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences. E-mail: jiayi.he@ia.ac.cn

Jiangyan Yi is with the Department of Automation, Tsinghua University, Bejing, China. E-mail: yijy@tsinghua.edu.cn

Jianhua Tao is with the Department of Automation and Beijing National Research Center for Information Science and Technology, Tsinghua University, Bejing, China. E-mail: jhtao@tsinghua.edu.cn (*Corresponding author)

Siding Zeng is with University of Chinese Academy of Sciences and the State Key Laboratory of Multi-modal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences. E-mail: zengsiding2023@ia.ac.cn

Hao Gu is with the State Key Laboratory of Multi-modal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences. E-mail: guhao2022@ia.ac.cn

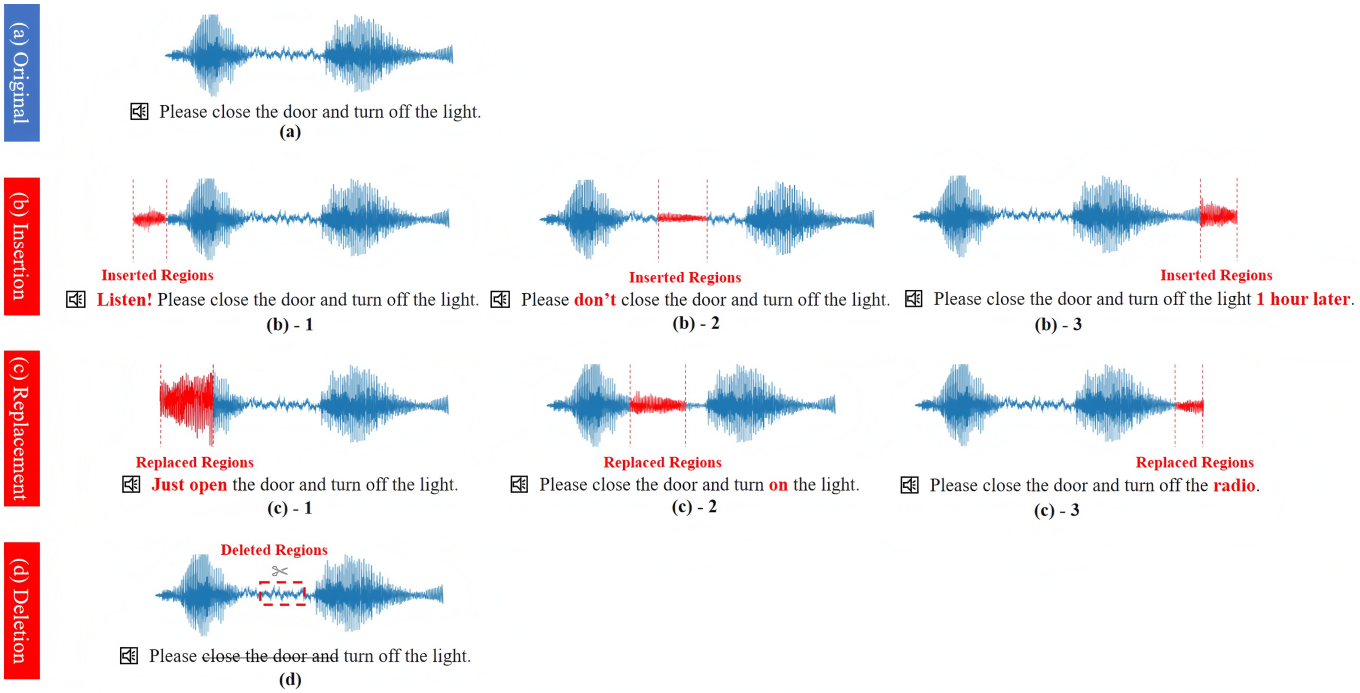


Fig. 2. Original bona fide audio and different types of partially deepfake audio: (a) Original bona fide audio; (b) Insertion at the beginning, middle, and the end of the utterance respectively; (c) Replacement at the beginning, middle, and the end of the utterance respectively; (d) Deletion at the middle of the utterance.

tasks.

- This survey provide a comprehensive summary, including the difficulty and specificity of the task, as well as the categories and the performance comparison of existing methods, which is particularly helpful for enhancing the understanding and raising community awareness.
- Specifically, the best performance for some commonly used diverse datasets are collected. Based on the current performance, the remaining challenges and limitations are discussed. At the meanwhile, some potential development trends are also discussed.

II. Definition and Characteristics of Partially Deepfake Audio Manipulation Localization

A. Definition of Partially Deepfake Audio

Partially deepfake audio refers to the audio with partial manipulations. The source of manipulated clips may be either synthetic or bona fide. The splicing method can include insertion, replacement, deletion, etc. Fig.2 shows the illustration of the different types. From the illustration, it can be seen that the meaning or tone of utterances can be changed by simply manipulating certain regions of it.

B. Differences from Fully Deepfake Audio

1) Process of Generation: The fully deepfake audio is usually entirely synthesized via text-to-speech (TTS)[36], voice conversion (VC)[36, 37], Emotion Fake[38], Scene Fake[39], etc., focusing on the fidelity and naturalness of the entire audio. While, for partially deepfake audio, only a few clips in a genuine audio are manipulated. The entire process generally includes three steps: selecting the splicing position, preparing the manipulated segments, and splicing the manipulated segments[13]. Its main concern is to ensure that the manipulated clips are highly consistent with real clips and avoiding leaving stitching marks. In summary, partially deepfake audio generation focuses on the high-quality operations of local substitution and seamless stitching, while fully deepfake audio generation emphasizes on global fidelity and naturalness.

2) Purpose of Spoofing: Fully deepfake audio spoofing is usually used to generate the voice of the target objects to achieve sound deception. While partially deepfake audio spoofing tends to change the expression of the original voice command by editing a few key words to implement the specific intent tampering.

C. Differences from Detection Tasks

Deepfake audio detection tasks mainly focus on the binary classification of genuine audio and fake or partially fake audio, and provides sentence-level labels, focusing more on the absolute authenticity of global features. While partially deepfake audio manipulation localization tasks emphasize to identify the manipulated regions in the audio and provide segment-level identification by discovering the local inconsistency of the audio itself (See Fig.1). In special circumstances, localization tasks can be considered as segment-level detection tasks when and only when the manipulated segment is fake.

III. Related Work

A. Fundamental Facilities

1) Datasets: To date, there have been several established datasets for partially deepfake audio localization tasks. The

TABLE I
THE DETAILS OF DATASETS FOR PARTIALLY DEEPFAKE AUDIO MANIPULATED REGIONS LOCALIZATION TASKS.

Ref.	Year	Dataset	Language	Modality	The Type of Manipulated Clips	Manipulation Methods	Access	#Utterances		
								Real	Fully Fake	Partially Fake
[12, 40]	2021	PartialSpoof	English	Audio	Generated	TTS/VC	Public	12,483	0	108,978
[13]	2021	HAD	Chinese	Audio	Generated	TTS	Public	53,612	53,612	53,612
[15]	2022	ADD2022Track2	Chinese	Audio	Generated/Real	TTS/Real	Public	5,319	45,367	1,052
[14]	2022	Psynd	English	Audio	Generated	TTS	Restrict	-	-	-
[41]	2022	LAV-DF	Multilingual	Audio/Video	Generated	TTS	Public	36,431	0	99,873
[16]	2023	ADD2023Track2	Chinese	Audio	Generated/Real	TTS/Real	Public	55,467	1,618	63,831
[42]	2024	AV-Deepfake1M	Multilingual	Audio/Video	Generated	TTS	Public	286,721	0	860,039
[17]	2025	LlamaPartialSpoof	English	Audio	Generated	TTS	Public	10,573	33,461	32,194
[43]	2025	AV-Deepfake1M++	Multilingual	Audio/Video	Generated	TTS	Restrict	-	-	-

infomaton is shown in Table I.

- **PartialSpoof Dataset**[1][12, 40]. This is the first English dataset proposed to focus on partially deepfake audio. It is built based on ASVspoof 2019 LA database[44, 45] and provides segment labels for various temporal resolutions[46]. In this datasets, every partially deepfake audio is a mixture of genuine and fake clips. Segments randomly chosen from a genuine audio are replaced with spoofed one and vice versa. Both segment-level labels and sentence-level labels are provides. Segments and utterances containing one or more generated frames are labeled as *spoof*, otherwise *bona fide*.
- **Half-truth Dataset (HAD)**[2][13]. This is the first Chinese partially deepfake audio dataset, built based on AISHELL-3 corpus[47], consisting of partially fake, fully fake, and real audio. Compared to the PartialSpoof database, instead of randomly choosing segments to pollute the raw audio, semantic coherence and word boundaries are considered during the manipulation generation.
- **ADD2022Track2 Dataset**[3][15]. It is designed to support the first Audio Deep synthesis Detection challenge (ADD 2022), consisting of partially fake, fully fake, and real audio. In this dataset, the partially fake audio is collected as an adaptation set, generating by manipulated the original genuine audio with real or synthesized clips. Test set consists of unseen genuine and partially fake audio, where some utterances are selected from Mandarin corpus AISHELL-1[48], AISHELL-3[47], and AISHELL-4[49].
- **ADD2023Track2 Dataset**[4][16]. It is designed to support the second Audio Deep synthesis Detection challenge (ADD 2023), consisting of partially fake, fully fake, and real audio. Similar to ADD2022Track2 dataset, the partially fake audio is also generating by manipulated the original genuine audio with either real or synthesized clips. The training and dev sets are also collected based on AISHELL-3. The test set includes unseen partially fake and real utterances. Different from ADD2022Track2 dataset, the training and dev sets consist of all of the three types. Besides, in test set, additional noise and format conversions were done, which significantly increased the difficulty in localization.
- **Partial Synthetic Detection dataset (Psynd)**[5][14]. This dataset consists of approximately 13 hours multi-speaker English corpus, based on LibriTTS[50], and the fake segments are injected into real utterances.
- **Localized Audio Visual DeepFake Dataset(LAV-DF)**[6][41]. This is the first large audio-visual deepfake dataset in manipulation localization tasks. The manipulation is rule-based and content-driven. The manipulation strategy is to replace strategic words with their antonyms, which leads to a significant change in the sentiment of the statement. In this dataset, the audio is extracted from video, and the real videos are sourced from the VoxCeleb2 dataset[51]. The partial fake is triggered by transcript manipulation, and the corresponding partially fake audio is generated by SV2TTS[52].
- **AV-Deepfake1M Dataset**[7][42]. It is a further step of content-driven audio-visual deepfake dataset for manipulation localization tasks. Different from LAV-DF, it employed ChatGPT for altering the real transcripts, ensuring the diversity and context consistent. It includes two additional challenging manipulation strategies, deletion and insertion, more than replacement. Besides, VITS[53] and YourTTS[54] are employed to generated the fully fake and partially fake audio. Its size is nearly ten times that of LAV-DF. Recently, AV-Deepfake1M++ dataset is released[43], containing over 2 million samples
- **LlamaPartialSpoof Dataset**[8][17]. It is a content-driven deepfake dataset with audio only. This dataset is designed to enhance the quality and diversity of fully fake and partially fake utterances, built based on LibriTTS. Inspired

[1]PartialSpoof: https://zenodo.org/records/5766198
[2]HAD: https://zenodo.org/records/10377492
[3]ADD2022Track2 Train&Dev: https://zenodo.org/records/12188127
ADD2022Track2 Adaption: https://zenodo.org/records/12188083
ADD2022Track2 Eval: https://zenodo.org/records/12187997
[4]ADD2023Track2 Train&Dev: https://zenodo.org/records/12176530
ADD2023Track2 Eval: https://zenodo.org/records/12176904
[5]Psynd: https://scholarbank.nus.edu.sg/handle/10635/227398
[6]LAV-DF: https://huggingface.co/datasets/ControlNet/LAV-DF
[7]AV-Deepfake1M: https://huggingface.co/datasets/ControlNet/AV-Deepfake1M
[8]LlamaPartialSpoof: https://huggingface.co/datasets/HaoY0001/LlamaPartialSpoof

by AV-Deepfake1M, Llama-3-8B-Instruct is employed to automatically alter sentences. The difference is that, in this dataset, the model is asked to change the transcript via several prompts instead of generating a series of replace, delete, and insert operations , improving the quality of manipulated transcription. Five TTS models are adopted to generate the fully fake and partially fake audio. The partially fake audio in this dataset is concatenated by real and fake segments. Post-process is done for both bona fide and the fake utterances.

2) Evaluation metrics:

- **Segment-level EER**. Zhang *et al*[12] proposed to adapt utterance-level EER to segment-level EER at first, named as point-based EER. The definition is showing below:

$$FPR = \frac{FP}{FP + TN} \tag{1}$$

$$FNR = \frac{FN}{FN + TP} \tag{2}$$

where FP refers to the real segments that are incorrectly detected as fake, and TN refers to the real segments that are correctly detected as real, and FN refers to the fake segments that are wrongly detected as real, and TP refers to the fake segments that are correctly detected as fake. When FPR = FNR, the common value is EER, which is widely used in the binary classification tasks. Then, to correct the precisions that caused by some potential misclassified regions and relieve the impact of diverse resolution, they modified it to range-based EER[55].

$$EER = \frac{FPR(\tau) + FNR(\tau)}{2}, \tag{3}$$

where

$$FPR(\tau) = \frac{\sum_{i \in Hypo} \sum_{j \in Ref} \mathcal{I}(Pred < \tau)\mathcal{T}(r_i, r_j)}{Duration \quad of \quad Negtive \quad Label}, \tag{4}$$

$$FNR(\tau) = \frac{\sum_{i \in Hypo} \sum_{j \in Ref} \mathcal{I}(Pred \geq \tau)\mathcal{T}(r_i, r_j)}{Duration \quad of \quad Positive \quad Label}. \tag{5}$$

$\mathcal{I}(\cdot)$ denotes the indicator function that outputs 1 when the condition is true and 0 otherwise. $\mathcal{T}(\cdot)$ records the overlap of two ranges. τ is obtain by binary search algorithm as described in [55]. Instead of labeling every segments, in this metric, reference labels are given according to the boundaries and duration for manipulation regions.

- **Segment-level Precision (P), Recall (R) and F1-score (F_1)**. Yi *et al*[13] proposed to use these metrics to evaluate the performance of localization accuracy, which are based on the duration of each segment.

$$P = \frac{TP}{TP + FN} \tag{6}$$

$$R = \frac{FN}{TP + FP} \tag{7}$$

$$F_1 = \frac{2PR}{P + R} \tag{8}$$

where TP refers to the fake segments that are correctly detected as fake, and FN refers to the fake segments that are incorrectly detected as real, and FP refers to the real segments that are wrongly detected as fake.

- **The weighted sum of sentence-level Accuracy(Acc) and segment-level** F_1. In ADD 2023 Track 2[16], the evaluation is designed to focus on both sentence-level and segment-level performance at the same time. Thus, it is defined as a weighted sum of *Sentence Accuracy* and *Segment* F_1, as shown in Eq.9.

$$Score = 0.3 \times Acc + 0.7 \times F_1, \tag{9}$$

where

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \tag{10}$$

$$F_1 = \frac{2PR}{P + R} \tag{11}$$

where the definition of TP, FP, TN, and FN are consistent with above. It is worth noting that these statistics in the Acc related formulas are at the sentence level, while those related to F_1 are at the segment level.

- **1D-Intersection over Union (IoU)**. Zhang *et al.* adopted 1D-Intersection over Union (IoU) as partially-spoofed audio detection evaluation[14]. The intersection indicates the number of segments that are correctly predicted. The union is the sum of intersection and twice the number of segments that are mispredicted, as shown in Eq. 12.

$$IoU = \frac{TP + TN}{TP + TN + 2 \times (FP + FN)} \tag{12}$$

The system will be considered as a good detector if $IoU > \frac{1}{3}$.

- **Average precision (AP) and average recall (AR)**. AP measures the performance by averaging precision at different recall levels, providing a comprehensive assessment of precision and recall.

$$AP = \sum_t (R_t - R_{t-1}) P_t \tag{13}$$

where R_t and P_t are the recall and precision at the threshold *t*. Usually, the threshold values are set at 0.5, 0.75, 0.9, 0.95.

AR focuses on the recall ability at different confidence thresholds, particularly useful in scenarios where high recall is essential, and average number of proposals *N* are usually set to 5, 10, 20, 50, 100.

$$AR = \frac{1}{N} \sum_{i=1}^{N} R(i) \tag{14}$$

B. Competitions

In order to prosper this developing topic, some competitions have been organized to facilitate technical communication, summarizing in Table II.

The first Audio Deep Synthesis Detection Challenge (ADD2022)[9] is held in 2022, organized by Jianhua Tao and

[9]ADD 2022: http://addchallenge.cn/add2022

TABLE II
THE EXISTING COMPETITIONS FOR PARTIALLY DEEPFAKE AUDIO TASK.(A: AUDIO, V: VIDEO)

Competitions	Track	Year	Mod.	Language	URL
ADD2022	Partially fake audio detection	2022	A	Chinese	http://addchallenge.cn/add2022
ADD2023	Manipulation region location	2023	A	Chinese	http://addchallenge.cn/add2023
2024 1M-Deepfakes Detection Challenge	Deepfake Temporal Localization	2024	AV	Multilingual	https://deepfakes1m.github.io/2024
2025 1M-Deepfakes Detection Challenge	Deepfake Temporal Localization	2025	AV	Multilingual	https://deepfakes1m.github.io/2025

Haizhou Li[15]. In this challenge, partially fake audio detection(PF) is firstly launched as an independent track, focusing on binary real/fake classification. Different from fully anti-spoofing task, it emphases on detecting the partially fake utterance with real or synthesized audio inserted from bona fide audio. EER is employed as the evaluation metric. The ADD 2022 is also launched as a Signal Processing Grand Challenge at the IEEE International Conference on Acoustics, Speech and Signal Processing in 2022 (ICASSP 2022). Additionally, based on this challenge, Tao *et al.* also initiated a workshop on Deepfake Detection for Audio Multimedia at ACM Multimedia 2022 (DDAM 2022)[10].

In 2023, the second Audio Deep Synthesis Detection Challenge (ADD 2023) is launched[11]. Different from ADD 2022, the setting of ADD 2023 goes beyond the goal of binary real/fake classification for entire utterances, which is the first competition focusing on localizing the manipulated intervals in partially fake audio (Track 2). The weighted sum of sentence-level Accuracy and segment-level F_1 is employed as the evaluation metric. The ADD 2023 challenge is also organized as part of the IJCAI 2023 Competitions and Challenges track, and the IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023) is organized based on it, leading widespread discussion within the scope. The systems on the leaderboard has become an important baseline for the following research in partially deepfake audio manipulation regions localization tasks.

The first 1M-Deepfakes detection challenge was launched in 2024, held by Abhinav Dhall *et al.* at ACM Multimedia 2024[56]. In this challenge, the *Task2: Deepfake Temporal Localization* is to find out the timestamps [start, end] in which the manipulation is done, aiming at multi-modal data. AV-Deepfake1M Dataset is released in this challenge, which is multi-modality and multilingual. Recently, the second 1M-Deepfakes detection challenge is in progress. The new challenge is based on AV-Deepfake1M++ dataset containing over 2 million samples[43].

IV. BRANCHES OF METHODS

To date, there have been several studies working on partially deepfake audio manipulation region localization. All of these methods could be divided into FOUR types as shown in Fig.3, and their strengths and weakness are summarized in Table III.

[10]DDAM 2022: http://addchallenge.cn/ddam2022
[11]ADD 2023: http://addchallenge.cn/add2023

A. Methods Based on Frame-level Authenticity

For this category, the manipulation regions are detected based on the authenticity of segments. Due to the fact that commonly used datasets employ fake segments as the splicing clips, such as PartialSpoof, most existing methods belong to this type. Usually, two-stage frameworks are designed, consisting of a front-end feature extractor and a back-end classifier. MFCC[61], LFCC[62], CQCC[63], Wav2vec[64, 65] and WavLM[66] are commonly employed as feature extractors while light convolutional neural network (LCNN)[67], ResNet[68], SENet[69] and long short-term memory (LSTM)[70] are representative classifiers.

For example, Zhang *et al.*[14] propose to use CQCC and ANN as feature extractors and classifier respectively. A post-processing is employed to filter extreme short fake or real segments to modify the results. Zhang *et al.*[24] introduce a binary-branch multi-task models by integrating squeeze-and-excitation (SE) blocks with LCNN (SELCNN) and a BiLSTM to implement the basic model, employing LFCC as a front-end feature extractor. Zhu *et al.*[32] add self-attention mechanism between SELCNN and BiLSTM to enhance the segment features, greatly improving the partially deepfake detection performance. Li *et al.*[28] adopt convolutional recurrent neural network (CRNN) to capture high temporal features and the context information. Li *et al.*[31] combine AASIST and Wav2Vec2 subsystems through multi-grained backend fusion to find out fake utterances or frames, where AASIST extracts features from utterance-level while Wav2Vec2 from segment-level. Martín-Doñas *et al.*[25] integrate Wav2Vec2 based feature extractor and BiLSTM to cluster the manipulated frames for partially deepfake detection.

Besides, some methods propose distinctive functional modules by combining these fundamental modules to enhance the performance. Xie *et al.*[19] propose temporal deepfake location (TDL) method to locate the manipulated regions. They devise an embedding similarity module to segregate authentic and synthetic frames within the embedding space to enhance the identification of genuine and fake distinctions at the embedding level. The result shows that it could achieve the EER at 7.04% on PartialSpoof dataset at 160ms resolution, which was once the best performance of this dataset. Besides, it is also demonstrated that it could achieve the EER of 11.23% on LAV-DF dataset, which reveals its good generalization ability. Inspired by this method, Dragar *et al.*[60] modified TDL to a window-based method, named as W-TDL, and combined it with the EVA visual transformer to identify and localize manipulated segments in audio and visual data, achieving the best performance on AV-Deepfake1M dataset.

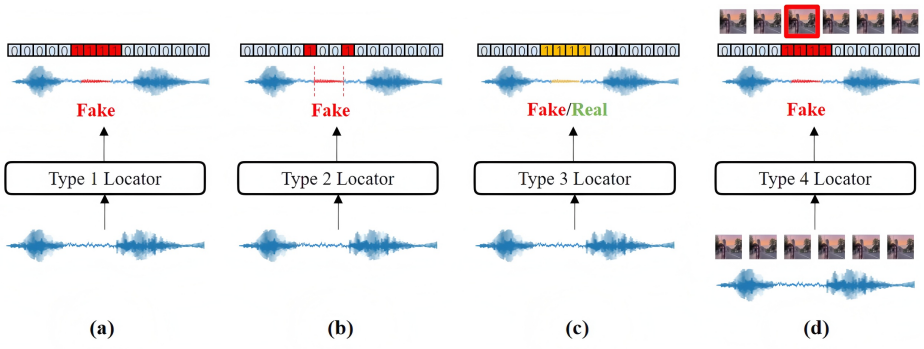


Fig. 3. Four categories of locators in existing studies. (a) Locator based on frame-level authenticity; (b) Locator based on boundary perception; (c) Locator based on frame-level inconsistency; (d) Locator based on multi-modality fusion.

TABLE III
THE STRENGTHS AND WEAKNESS FOR EACH TYPE OF METHOD.

Type	Properties	Strengths	Weakness	Related methods
1	Frame-level Authenticity	It is straightforward and constitutes the majority in existing research.	It may fail to locate the manipulation regions when the splicing clips are bona fide	SPF[26], TDL[19]
2	Boundary Perception	It focuses on detecting stitching traces to avoid relying entirely on frame-level authenticity.	It may fail when the splicing boundaries are hidden intentionally	CFPRF[57], BAM[35]
3	Frame-level Inconsistency	It focuses on the inconsistency between frames instead of the authenticity, overcoming the weakness of the former two types.	For long lasting audio, the information in the utterances may change, the effectiveness needs further validation	PET[33], AGO[34], GNCL[58]
4	Multi-Modality Fusion	It integrates multimodal forgery information and represents a new trend in recent research.	It focuses more on the visual modality, and further explorations are needed in audio modality.	UMMAFormer[59], W-TDL[60]

Cai *et al.*[21, 26, 71] designed the anti-spoofing detection system (SPF) to detect the fake frames embedding in the real audio. The Wav2Vec and WavLM are employed as feature extractors, and ResBlock is further used to learning the feature in-depth. Finally, transformer encoder and bidirectional long short term memory network (BiLSTM) are adopted as backend classifiers. It demonstrated that SPF achieves the champion of ADD 2023 Track 2 with the score of 0.6713.

Additionally, MUSAN noise[72], reverberations and some other data augmentation strategies are usually employed to help enhance the robust of performance[20, 27, 28, 71]. Multi-domain feature fusion strategy has also been proposed[73].

However, although this type of methods dominates currently, it may fail to locate the manipulation regions when the splicing clips are bona fide.

B. Methods Based on Boundary Perception

For this category, the manipulation regions are detected via splicing traces. The intention is to focus on stitching traces and avoid relying entirely on fragment-level authenticity. The study shows that the partially spoofed audio-trained CMs significantly focus on the transition regions created by the overlap-add operation during the dataset creation[74]. However, the biggest obstacle encountered by such methods is data bias. Thus, some of these existing methods appear simultaneously with frame-level authenticity classifiers.

Wu *et al.*[22] introduce a question-answering (QA) strategy based on SE-ResNet architecture with self-attention mechanism to locate the manipulated regions by predicting the start and end positions of clips. Zeng *et al.*[18] adopt a ResNet-based model for splicing traces localization, both time and frequency features are considered. The localization probability are obtained via 4 consecutive frames. They conducted experiments on their own dataset and achieved the F_1 at 0.741 in the test set with chunk size of 64 frames. Cai *et al.*[21, 26, 71] proposed a dual-head system to detect the a frame-level anti-spoofing and locate the boundary simultaneously. Boundary detection system (BDR) is designed to perceive the splicing boundaries with the same network structure as SPF. It is claimed to achieve the frame-level EER at 0.064% on ADD2023 Track 2 dev set with training on ADD2023 Track 2 train set and at 1.74% on PartialSpoof eval set with training on PartialSpoof train set for boundary frames detection. Wu *et al.*[57] introduced a coarse-to-fine proposal refinement framework (CFPRF) to locate the partially fake. They proposed the temporal forgery localization (TFL) network to predict the precise timestamps at which these forgery segments start and end. It reveals that CFPRF could achieve the EER at 0.08 on HAD dataset, 7.41 on PartialSpoof dataset, and 0.82 on LAV-DF dataset for fake segments localization, which are claimed to be superior to the method mentioned in Ref.[71] and [46]. Zhong *et al.*[35] proposed boundary-aware attention Mechanism (BAM), consisting of boundary enhancement (BE) module and boundary frame-wise attention (BFA) module, to improve the accuracy and localization capability by using boundary information. BE aims to extract intra-frame and inter-frame information to enhance boundary features for splicing boundary detection and authenticity detection. BFA aims to use boundary prediction results to explicitly control

the feature interaction between frames, in order to effectively distinguish between real and fake frames. When using WavLM as the front-end feature, the BAM method ontained an EER of 3.58% at a resolution of 160ms, achieving the state-of-the-art performance on the PartialSpoof dataset.

Obviously, the existing methods inherit the drawbacks of the first type of method if they partially relied on the frame-level authenticity. Besides, facing with the increasingly advanced splicing technology, in real adversarial attack scenarios, the splicing boundaries will be intentionally hidden, and such methods are prone to failure.

C. Methods Based on Frame-level Inconsistency

For this category, the manipulation regions are detected based on the inconsistency between the manipulated regions and non-manipulated ones, which can provide effective solutions to overcome the difficulties encountered by the two types mentioned above. Existing studies indicate that the systems composed of frame-wise consistency related modules usually exhibit superior performance. According to the existing methods, there are three subtypes:

1) Difference-Aware Between Real and Fake Frames: In CFPRF[57], Difference-Aware Feature Learning Module (DAFL) is proposed to enhance the difference between real and fake frames. Also, In TDL[19], the embedding similarity module is designed to capture the differences in feature learning between real and fake frames and employed as mask to enhance the diversity. The designation of BE module in BAM[35] is also the same.

2) Distribution Shift Between Manipulated and Non-manipulated Regions: Zeng *et al.*[34] proposed the adversarial training and gradient optimization (AGO) method to locate the partially deepfake segments by focusing on the distribution shift between manipulated and non-manipulated regions, which provides a new perspective to address the issue. Gradient reversal layer (GRL) is employed to reduce the dependence of model on specific domain features and enhance the generalization ability. The results show that AGO could achieve the segment-level F_1 score at 0.7187 and ADD2023 score at 0.8254 on ADD2023 Track 2 dataset, which is a relative improvement of 22.82% than SPF[21, 26, 71] without any data augmentation strategies. In PartialSpoof dataset, it could achieve the EER at 6.79, which is superior than that of CFPRF[57] at 7.41.

3) Inconsistency Between Manipulated and Non-manipulated Regions: He *et al.*[33] initialed a partially deepfake audio localization method via empirical wavelet transform and temporal self-consistency learning (PET), locating manipulated regions via temporal self-consistency learning of high-frequency components. Different from existing methods, PET directly utilizes the frame-wise similarity of high-frequency components as a feature to capture the inconsistency among frames. It is a location-only system that could achieve the state-of-the-art segment-level F_1 score at 0.7397, 2.92% relative improvement compared to AGO[34] and 21.94% higher than that of SPF[21, 26, 71] at 0.6066. Ge *et al.*[58] proposed a graph neural network with consistency loss (GNCL) to locate the spoofed segments. The consistency-enhanced loss function is introduced to bridge different. It achieves the EER at 11.81% on PartialSpoof dataset at a 20ms resolution.

However, although these methods have achieved good results for audio clips with a few seconds long, their effectiveness for longer lasting audio, such as continuous recordings spanning several hours or more, needs further validation.

D. Methods Based on Multi-Modality Fusion

AVFusion[76] is the first model to jointly consider audio and video modalities for temporal action localization, aiming to locate the start and end timestamps of activities in the video stream. Based on that, some studies[41, 78] are initialed for temporal forgery localization to locate the start and end timestamps of manipulated segments. Cai *et al.*[41] proposed BA-TFD and BA-TFD+, two multi-modality methods, for content-driven partially deepfake audio-video detection and illustrated its effectiveness on LAV-DF dataset. They are now also considered as baseline methods on the LAV-DF dataset. Zhang *et al.*[59] proposed UMMAFormer to predict forgery segments and their corresponding start and end timestamps in untrimmed videos or audios, considering three scenarios: visual-only, audio-only, and joint audio-visual modalities. In UMMAFormer, a Temporal Feature Abnormal Attention (TFAA) module is built from reconstruction learning and Cross-Reconstruction Attention Transformer (CRATrans) block to identify abnormal segments. The results reported that, compared to BA-TFD, the AP@0.5 has increased from 76.90% to 98.83%, and from 0.29% to 37.61% at AP@0.95 on LAV-DF dataset. Further more, CFPRF[57] refreshed the AP@0.95 to 88.61%. Besides, Cai *et al.* further expanded the LAV-DF dataset to the AV-Deepfake1M dataset[42], and organized the 2024 1M-Deepfakes Detection Challenge on ACM Multimedia 2024[12]. In the challenge, W-TDL, proposed by Dragar *et al.*[60], is confirmed to outperform existing state-of-the-art techniques on the AV-Deepfake1M dataset.

However, currently this type of method focuses more on the aspect of visual modality, and further explorations are needed in audio modality.

Additionally, beyond these types, the audio copy-move forgery detection task also can be considered as a simplified situation of this manipulation regions localization tasks, which is a legacy technique with the copied frames being selected from the audio itself and then being inserted or replaced at certain position in the audio[79]. The pipeline of solving audio copy-move forgery detection usually begins with voice activity detection module(VAD). Then discrete cosine transform(DCT) coefficients, the constant Q spectral sketches (CQSS), discrete Fourier transform (DFT), MFCC, etc. are employed as feature extractor to obtain the feature representation for segments. Euclidean distance(ED), dynamic time warping(DTW) and cosine similarity are adopted as measurement to calculate pairwise distance or similarity between segments in order to locate the copy-move forgery regions[80–89]. Obviously,

[12]2024 1M-Deepfakes Detection Challenge: https://deepfakes1m.github.io/2024

TABLE IV
COMPARISON OF PROMINENT METHODS IN PARTIALLY DEEPFAKE AUDIO LOCALIZATION TASKS. THE METRICS MENTIONED IN THE TABLE ARE ALL IN SEGMENT-LEVEL. THE RESULTS ARE ALL FROM CITATIONS. (THE "*" INDICATES THAT THE CFPRF METHOD WAS TRAINED ON THE TRAINING SET OF THE LAV-DF DATASET AND TESTED ON ITS TEST SET, WHILE THE RESULTS OF OTHER METHODS ON LAV-DF ARE TRAINED WITH PARTIALSPOOF TRAIN SET.)

Model	Year	PartialSpoof		ADD2023 Track 2		LAV-DF	
		EER(%)↓ /Resolution	F_1↑	EER(%)↓	F_1↑ /Resolution	EER(%)↓	F_1↑
LCNN-BLSTM(w LFCC)[12, 19]	2021	16.21 / 160ms	-	-	-	17.89	0.8338
LCNN-BLSTM(w W2V2-XLS-R)[12, 19]	2021	9.87 / 160ms	-	-	-	15.35	0.7650
SELCNN-BLSTM[24]	2021	15.93 / 160ms	-	-	-	-	-
SPF(w WavLM)[26, 71]	2023	-	0.9296	-	0.6066 / 20ms	-	-
TranssionADD[20, 75]	2023	-	-	-	0.5460 / 160ms	-	-
CRNN[28]	2023	-	-	-	0.5449 / 10ms	-	-
Multi-grained Backend Fusion[31]	2023	-	-	-	0.5253 / 20ms	-	-
Vicomtech[25]	2023	-	-	-	0.5167 / 20ms	-	-
TDL[19]	2024	7.04 / 160ms	-	-	-	**11.23**	**0.8551**
BAM(w WavLM-Large)[35]	2024	**3.58 / 160ms**	**0.9609**	-	-	-	-
CFPRF[57]	2024	7.41 / Not found	0.9389	-	-	**0.82***	**0.9956***
AGO[34]	2025	6.79 / 40ms	0.9436	-	0.7187 / 40ms	-	-
GNCL[58]	2025	11.81 / 20ms	0.8979	-	-	-	-
PET[33]	2025	-	-	29.50	**0.7397 / 10ms**	-	-

TABLE V
COMPARISON OF PROMINENT METHODS IN PARTIALLY DEEPFAKE AUDIO-VIDEO LOCALIZATION TASKS. THE RESULTS ARE ALL FROM CITATIONS.

Model	Dataset	Year	Mod.	AP@0.95	AP@0.9	AP@0.75	AP@0.5	AR@5	AR@10	AR@20	AR@50
AVFusion[76]	LAV-DF	2021	AV	0.11	-	23.89	65.38	-	62.98	59.26	54.80
BA-TFD[41]		2022	AV	0.29	-	47.06	76.90	-	59.32	61.19	64.52
BA-TFD+[60, 77]		2023	AV	4.44	-	84.96	96.30	-	78.75	79.40	80.48
UMMAFormer[59, 60]		2023	AV	37.61	-	**95.54**	**98.83**	-	92.10	92.42	**92.48**
CFPRF[57]		2024	A	**88.64**	**91.65**	93.47	94.52	**93.51**	**93.51**	**93.51**	-
BA-TFD[41]	AV-Deepfake1M	2022	AV	0.02	0.19	6.34	37.37	26.82	30.66	35.95	45.55
BA-TFD+[60, 77]		2023	AV	0.03	0.48	13.64	44.42	29.88	34.67	40.37	48.86
UMMAFormer[59, 60]		2023	AV	1.58	07.65	28.07	51.64	40.27	42.09	43.45	44.07
W-TDL[60]		2024	AV	**50.66**	**70.43**	**88.75**	**94.75**	**88.78**	**89.13**	**89.17**	**89.17**

methods for audio copy-move forgery detection are mainly based on the pairwise similarity of waveform or spectrum between clips, which will have limitations when used for partially deepfake manipulation regions localization generated via cutting-edge techniques. But they can inspire us to achieve the goal by constructing deeper feature for localization via frame-wise similarities.

V. COMPARISONS OF EXISTING METHODS

In this section, the comparisons of some methods that utilize common datasets are demonstrated, including PartialSpoof, ADD2023 Track 2 dataset, LAV-DF and AV-Deepfake1M dataset. There are mainly two groups for comparison, one for audio-only (See Table IV) and another for audio-video datasets (See Table V).

In Table IV, it reveals the significant technological improvements in partially deepfake audio localization. The results show that, in the past five years since the issue was raised, the segment-level EER of the PartialSpoof dataset has decreased from 16.21% to 3.58%(BAM), and the segment-level F_1 has increased to 0.9609. For ADD2023 Track 2 dataset, the segment-level F_1 has increased from 0.6066 to 0.7397(PET). CFPRF, as a uni-modality model, could achieve the segment-level EER at 0.82%, demonstrating the potential ability of partially deepfake audio localization methods in multi-modality partially deepfake datasets. Some methods also show well cross-domain localization capabilities. Specifically, TDL could achieve the segment-level EER at 11.23% on LAV-DF dataset while training on the PartialSpoof train set.

Table V shows the comparison of partially deepfake localization models for audio-video datasets, illustrating remarkable progress in multi-modality partially deepfake localization tasks that incorporating audio as one of the key modalities. The latest research shows that, for the LAV-DF dataset, the AP@0.95 has soared from 0.29(BA-TFD) to 88.64(CFPRF) and the AP@0.5 has increased from 76.90(BA-TFD) to 98.83(UMMAFormer). The scores of each item for average recall(AR) have approximate 50% relative improvements. For AV-Deepfake1M, the AP@0.95 has soared from 0.02(BA-TFD) to 50.66(W-TDL) and the AP@0.5 has increased from 37.37(BA-TFD) to 94.97(W-TDL). The score of AR@5 surges from 26.82(BA-TFD) to 88.78(W-TDL) and AR@50 has achieved a relative improvement of approximately 96%. Besides, the superior performance of CFPRF and W-TDL emphasizes the potential benefits of partially deepfake audio localization methods in multi-modality partially deep-

fake localization tasks that incorporating audio as one of the key modalities.

VI. Challenges and Development Trends

A. Challenges and Limitations

1) ***Insufficient localization accuracy****:* According to Sec.V, although some significant improvements have achieved, however, as shown in Table V, there is still a long way to go for lower resolution and complex situations, such as audio with "truth for truth" manipulation, smoothing processing for splicing boundaries, etc.. Besides, the distribution shift for manipulated clips and environment shift for long lasting audio are also needed to be taken into considerations.

2) ***Lack of evidence to support the results****:* The existing methods typically use binary sequences to indicate the location of the manipulations, answering the question of 'what' but lacking a response to 'why'. Specifically, in practical applications, more detailed physical evidence is needed to support the results, such as some indicators or measurements refer to spectral discontinuities, changes in timbre, phase information missing, etc.

B. Potential Trends

1) ***Focusing on the inconsistency between manipulated and non-manipulated regions instead of their authenticity****:* Manipulations with real clips is very likely to occur in real-world scenarios, so in order to improve the practical ability and enhance the generalization capability, researchers need to shift their attentions to the essential features that are highly relevant to the differences between manipulated and non-manipulated regions, such as acoustic feature distribution shifting, noise inconsistency, sound field inconsistency, emotional inconsistency, etc..

2) ***Utilizing LLM-based methods to provide the evidence****:* Regarding the issue of lacking physical evidence, LLM may be helpful. There have been some studies on the application of speech LLM, but there has been no breakthrough in partially deepfake audio localization and its forensics. Researchers may further construct indicators related to the physical evidence and take full advantage of LLM's reasoning capabilities to obtain the physical evidence.

3) ***Expanding to multi-modality deepfake localization tasks****:* Given the newly proposed challenge competitions and datasets, the trend of multi-modal partially deepfake localization has emerged, but according to the studies, the current multi-modal deepfake localization evaluation usually focuses on visual modality, maybe in the near future, the audio part will play an important role.

VII. Conclusions

In this survey, we sort out the route to development of partially deepfake localization tasks, including datasets, evaluation metrics, challenge competitions, branches of existing methods, current limitations and potential trends, providing a comprehensive insight of this scope for beginners to catch up with. Specifically, we elaborate on the definition of partially deepfake audio localization and sorted out the current research status, including the method with the best performance on diverse datasets. Based on the achievements already made, potential trends of this scope is discussed. We hope this survey could be the reference for later researchers and bring about deeper thought and exploration.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) (No.62206278, No. 62322120, No. 62306316).

References

[1] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6367–6371.

[2] J. Xue, C. Fan, J. Yi, C. Wang, Z. Wen, D. Zhang, and Z. Lv, "Learning from yourself: A self-distillation method for fake speech detection," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[3] S. Ding, Y. Zhang, and Z. Duan, "Samo: Speaker attractor multi-center one-class learning for voice anti-spoofing," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[4] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639314000788

[5] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, "Advances in anti-spoofing: from the perspective of asvspoof challenges," *APSIPA Transactions on Signal and Information Processing*, vol. 9, p. e2, 2020.

[6] C. B. Tan, M. H. A. Hijazi, N. Khamis, P. N. E. B. Nohuddin, Z. Zainol, F. Coenen, and A. Gani, "A survey on presentation attack detection for automatic speaker verification systems: State-of-the-art, taxonomy, issues and future direction," *Multimedia Tools and Applications*, vol. 80, no. 21, pp. 32 725–32 762, 2021.

[7] A. Mittal and M. Dua, "Automatic speaker verification systems and spoof detection techniques: review and analysis," *International Journal of Speech Technology*, vol. 25, no. 1, pp. 105–134, 2022.

[8] Z. Almutairi and H. Elgibreen, "A review of modern audio deepfake detection methods: Challenges and future directions," *Algorithms*, vol. 15, no. 5, p. 155, 2022.

[9] J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, and Y. Zhao, "Audio deepfake detection: A survey," 2023.

[10] Y. Xie, Y. Lu, R. Fu, Z. Wen, Z. Wang, J. Tao, X. Qi, X. Wang, Y. Liu, H. Cheng, L. Ye, and Y. Sun,

"The codecfake dataset and countermeasures for the universally detection of deepfake audio," 2024. [Online]. Available: https://arxiv.org/abs/2405.04880

[11] A. Alali and G. Theodorakopoulos, "Partial fake speech attacks in the real world using deepfake audio," *Journal of Cybersecurity and Privacy*, vol. 5, no. 1, 2025.

[12] L. Zhang, X. Wang, E. Cooper, J. Yamagishi, J. Patino, and N. W. D. Evans, "An initial investigation for detecting partially spoofed audio," *ArXiv*, vol. abs/2104.02518, 2021.

[13] J. Yi, Y. Bai, J. Tao, H. Ma, Z. Tian, C. Wang, T. Wang, and R. Fu, "Half-Truth: A Partially Fake Audio Detection Dataset," in *Interspeech 2021*, 2021, pp. 1654–1658.

[14] B. Zhang and T. Sim, "Localizing fake segments in speech," in *2022 26th International Conference on Pattern Recognition (ICPR)*, 2022, pp. 3224–3230.

[15] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, S. Liang, S. Wang, S. Zhang, X. Yan, L. Xu, Z. Wen, and H. Li, "Add 2022: the first audio deep synthesis detection challenge." in *ICASSP*. IEEE, 2022.

[16] J. Yi, J. Tao, R. Fu, X. Yan, C. Wang, T. Wang, C. Y. Zhang, X. Zhang, Y. Zhao, Y. Ren, L. Xu, J. Zhou, H. Gu, Z. Wen, S. Liang, Z. Lian, S. Nie, and H. Li, "Add 2023: the second audio deepfake detection challenge," 2023.

[17] H.-T. Luong, H. Li, L. Zhang, K. A. Lee, and E. S. Chng, "Llamapartialspoof: An llm-driven fake speech dataset simulating disinformation generation," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[18] Z. Zeng and Z. Wu, "Audio splicing localization: Can we accurately locate the splicing tampering?" in *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2022, pp. 120–124.

[19] Y. Xie, H. Cheng, Y. Wang, and L. Ye, "An efficient temporary deepfake location approach based embeddings for partially spoofed audio detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 966–970.

[20] J. Liu, Z. Su, H. Huang, C. Wan, Q. Wang, J. Hong, B. Tang, and F. Zhu, "Transsionadd: A multi-frame reinforcement based sequence tagging model for audio deepfake detection," *arXiv preprint arXiv:2306.15212*, 2023.

[21] Z. Cai, W. Wang, and M. Li, "Waveform boundary detection for partially spoofed audio," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[22] H. Wu, H.-C. Kuo, N. Zheng, K.-H. Hung, H.-Y. Lee, Y. Tsao, H.-M. Wang, and H. Meng, "Partially fake audio detection by self-attention-based fake span discovery," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9236–9240.

[23] L. Wang, B. Yeoh, and J. W. Ng, "Synthetic voice detection and audio splicing detection using se-res2net-conformer architecture," *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 115–119, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252762161

[24] L. Zhang, X. Wang, E. Cooper, and J. Yamagishi, "Multi-task learning in utterance-level and segmental-level spoof detection," *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021.

[25] J. M. Martín-Doñas and A. Álvarez, "The vicomtech partial deepfake detection and location system for the 2023 add challenge," in *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis*, 2023.

[26] Z. Cai, W. Wang, Y. Wang, and M. Li, "The dku-dukeece system for the manipulation region location task of add 2023," *arXiv preprint arXiv:2308.10281*, 2023.

[27] J. Zhang, H. Liu, M. Deng, J. Wang, Y. Sun, L. Xu, and J. Li, *An Improved System for Partially Fake Audio Detection Using Pre-trained Model*, 02 2024, pp. 346–353.

[28] K. Li, X.-M. Zeng, J.-T. Zhang, and Y. Song, "Convolutional recurrent neural network and multitask learning for manipulation region location," in *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis*, vol. 750, 2023.

[29] M. H. Rahman, M. Graciarena, D. Castan, C. Cobo-Kroenke, M. McLaren, and A. Lawson, "Detecting synthetic speech manipulation in real audio recordings," in *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2022, pp. 1–6.

[30] A. K. Kumar, D. Paul, M. Pal, M. Sahidullah, and G. Saha, "Speech frame selection for spoofing detection with an application to partially spoofed audio-data," *International Journal of Speech Technology*, vol. 24, pp. 193–203, 2021.

[31] J. Li, L. Li, M. Luo, X. Wang, S. Qiao, and Y. Zhou, "Multi-grained backend fusion for manipulation region location of partially fake audio," in *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis*, vol. 755, 2023.

[32] Y. Zhu, Y. Chen, Z. Zhao, X. Liu, and J. Guo, "Local self-attention-based hybrid multiple instance learning for partial spoof speech detection," *ACM Transactions on Intelligent Systems and Technology*, vol. 14, no. 5, pp. 1–18, 2023.

[33] J. He, J. Yi, J. Tao, and S. Zeng, "Pet: High-frequency temporal self-consistency learning for partially deepfake audio localization," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[34] S. Zeng, J. Yi, J. Tao, J. He, Z. Lian, S. Liang, C. Zhang, Y. Chen, and X. Zhang, "Adversarial training and gradient optimization for partially deepfake audio localization," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[35] J. Zhong, B. Li, and J. Yi, "Enhancing partially spoofed audio localization with boundary-aware attention mechanism," in *Proc. Interspeech 2024*, 2024, pp. 4838–4842.

[36] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre,

and H. Li, "Spoofing and countermeasures for speaker verification," *Speech Communication*, vol. 66, no. C, p. 130–153, feb 2015. [Online]. Available: https://doi.org/10.1016/j.specom.2014.10.005

[37] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Commun.*, vol. 137, no. C, p. 1–18, feb 2022. [Online]. Available: https://doi.org/10.1016/j.specom.2021.11.006

[38] Y. Zhao, J. Yi, J. Tao, C. Wang, X. Zhang, and Y. Dong, "Emofake: An initial dataset for emotion fake audio detection," 2023.

[39] J. Yi, C. Wang, J. Tao, Z. Tian, C. Fan, H. Ma, and R. Fu, "Scenefake: An initial dataset and benchmarks for scene fake audio detection," *Pattern Recognit.*, vol. 152, p. 110468, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253499099

[40] L. Zhang, X. Wang, E. Cooper, J. Yamagishi, J. Patino, and N. Evans, "PartialSpoof Database - Partially Spoofed Audio Dataset for Anti-spoofing," 2022. [Online]. Available: https://doi.org/10.5281/zenodo.5766198

[41] Z. Cai, K. Stefanov, A. Dhall, and M. Hayat, "Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization," in *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2022, pp. 1–10.

[42] Z. Cai, S. Ghosh, A. P. Adatia, M. Hayat, A. Dhall, T. Gedeon, and K. Stefanov, "Av-deepfake1m: A large-scale llm-driven audio-visual deepfake dataset," in *Proceedings of the 32nd ACM International Conference on Multimedia*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 7414–7423. [Online]. Available: https://doi.org/10.1145/3664647.3680795

[43] A. Dhall, Z. Cai, S. Ghosh, K. Stefanov, M. Haris, and U. Tariq. (2025) 2025 1m-deepfakes detection challenge. Example Organization. Accessed on 2025-05-12. [Online]. Available: https://deepfakes1m.github.io/2025

[44] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. Le Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J.-F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang, Z.-H. Ling *et al.*, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230820300474

[45] M. Todisco, X. Wang, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *Proc. of Interspeech 2019*, 2019.

[46] L. Zhang, X. Wang, E. Cooper, N. Evans, and J. Yamagishi, "The partialspoof database and countermeasures for the detection of short fake speech segments embedded in an utterance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, p. 813–825, 2023. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2022.3233236

[47] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A Multi-Speaker Mandarin TTS Corpus," in *Proc. Interspeech 2021*, 2021, pp. 2756–2760.

[48] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017, pp. 1–5.

[49] Y. Fu, L. Cheng, S. Lv, Y. Jv, Y. Kong, Z. Chen, Y. Hu, L. Xie, J. Wu, H. Bu, X. Xu, J. Du, and J. Chen, "Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario," in *Interspeech 2021*, 08 2021, pp. 3665–3669.

[50] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," *arXiv e-prints*, p. arXiv:1904.02882, Apr. 2019.

[51] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *Interspeech 2018*, 2018.

[52] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in neural information processing systems*, vol. 31, 2018.

[53] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[54] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International conference on machine learning*. PMLR, 2022, pp. 2709–2720.

[55] L. Zhang, X. Wang, E. Cooper, N. Evans, and J. Yamagishi, "Range-Based Equal Error Rate for Spoof Localization," in *Proc. INTERSPEECH 2023*, 2023, pp. 3212–3216.

[56] Z. Cai, A. Dhall, S. Ghosh, M. Hayat, D. Kollias, K. Stefanov, and U. Tariq, "1m-deepfakes detection challenge," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 11 355–11 359.

[57] J. Wu, W. Lu, X. Luo, R. Yang, Q. Wang, and X. Cao, "Coarse-to-fine proposal refinement framework for audio temporal forgery detection and localization," in *Proceedings of the 32nd ACM International Conference on Multimedia*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 7395–7403.

[58] Z. Ge, X. Xu, H. Guo, Z. Yang, and B. Schuller, "Gncl: A graph neural network with consistency loss for segment-level spoofed speech detection," in *ICASSP 2025 - 2025*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025, pp. 1–5.

[59] R. Zhang, H. Wang, M. Du, H. Liu, Y. Zhou, and Q. Zeng, "Ummaformer: A universal multimodal-adaptive transformer framework for temporal forgery localization," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 8749–8759. [Online]. Available: https://doi.org/10.1145/3581783.3613767

[60] L. Dragar, P. Rot, P. Peer, V. Štruc, and B. Batagelj, "W-tdl: Window-based temporal deepfake localization," in *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, ser. MRAC '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 24–29. [Online]. Available: https://doi.org/10.1145/3689092.3689410

[61] L.-W. Chen, W. Guo, and L.-R. Dai, "Speaker verification against synthetic speech," in *2010 7th International Symposium on Chinese Spoken Language Processing*, 2010, pp. 309–312.

[62] M. Todisco, H. Delgado, K. A. Lee, M. Sahidullah, and J. Yamagishi, "Integrated presentation attack detection and automatic speaker verification: Common features and gaussian back-end fusion," in *INTERSPEECH 2018*, 2018.

[63] M. Todisco, H. Delgado, and N. W. D. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients," in *The Speaker and Language Recognition Workshop*, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:16551583

[64] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-training for Speech Recognition," *arXiv e-prints*, p. arXiv:1904.05862, Apr. 2019.

[65] H. Tak, M. Todisco, X. Wang, J. weon Jung, J. Yamagishi, and N. W. D. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," *ArXiv*, vol. abs/2202.12233, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247084242

[66] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, p. 1505–1518, Oct. 2022. [Online]. Available: http://dx.doi.org/10.1109/JSTSP.2022.3188113

[67] Z. Wu, R. K. Das, J. Yang, and H. Li, "Light convolutional neural network with feature genuinization for detection of synthetic speech attacks," in *Interspeech*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221819604

[68] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep Residual Neural Networks for Audio Spoofing Detection," *arXiv e-prints*, p. arXiv:1907.00501, Jun. 2019.

[69] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.

[70] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[71] Z. Cai and M. Li, "Integrating frame-level boundary detection and deepfake detection for locating manipulated regions in partially spoofed audio forgery attacks," *Comput. Speech Lang.*, vol. 85, no. C, Apr. 2024. [Online]. Available: https://doi.org/10.1016/j.csl.2023.101597

[72] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," *arXiv e-prints*, p. arXiv:1510.08484, Oct. 2015.

[73] A. K. Singh Yadav, K. Bhagtani, S. Baireddy, P. Bestagini, S. Tubaro, and E. J. Delp, "Mdrt: Multi-domain synthetic speech localization," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 171–11 175.

[74] T. Liu, L. Zhang, R. K. Das, Y. Ma, R. Tao, and H. Li, "How do neural spoofing countermeasures detect partially spoofed audio?" in *Proc. Interspeech 2024*, 2024, pp. 1105–1109.

[75] J. Yi, C. Y. Zhang, J. Tao, C. Wang, X. Yan, Y. Ren, H. Gu, and J. Zhou, "Add 2023: Towards audio deepfake detection and analysis in the wild," 2024. [Online]. Available: https://arxiv.org/abs/2408.04967

[76] A. Bagchi, J. Mahmood, D. Fernandes, and R. K. Sarvadevabhatla, "Hear me out: Fusional approaches for audio augmented temporal action localization," *arXiv preprint arXiv:2106.14118*, 2021.

[77] Z. Cai, S. Ghosh, A. Dhall, T. Gedeon, K. Stefanov, and M. Hayat, "Glitch in the matrix: A large scale benchmark for content driven audio–visual forgery detection and localization," *Computer Vision and Image Understanding*, vol. 236, p. 103818, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314223001984

[78] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian, "Not made for each other- audio-visual dissonance-based deepfake detection and localization," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 439–447. [Online]. Available: https://doi.org/10.1145/3394171.3413700

[79] M. Imran, Z. Ali, S. T. Bakhsh, and S. Akram, "Blind detection of copy-move forgery in digital audio forensics," *IEEE Access*, vol. 5, pp. 12 843–12 855, 2017.

[80] Z. Su, M. Li, G. Zhang, Q. Wu, M. Li, W. Zhang, and X. Yao, "Robust audio copy-move forgery detection using constant q spectral sketches and ga-svm," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 5, pp. 4016–4031, 2023.

[81] C. Li, Y. Sun, X. Meng, and L. Tian, "Homologous audio copy-move tampering detection method based on pitch," in *2019 IEEE 19th International Conference on*

Communication Technology (ICCT). IEEE, 2019, pp. 530–534.

[82] F. Akdeniz and Y. Becerikli, "Linear prediction coefficients based copy-move forgery detection in audio signal," in *2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. IEEE, 2022, pp. 770–773.

[83] B. Ustubioglu, B. Küçükuğurlu, and G. Ulutas, "Robust copy-move detection in digital audio forensics based on pitch and modified discrete cosine transform," *Multimedia Tools and Applications*, vol. 81, no. 19, pp. 27 149–27 185, 2022.

[84] F. Akdeniz and Y. Becerikli, "Detecting audio copy-move forgery with an artificial neural network," *Signal, Image and Video Processing*, vol. 18, no. 3, pp. 2117–2133, 2024.

[85] B. Ustubioglu, G. Tahaoglu, and G. Ulutas, "Detection of audio copy-move-forgery with novel feature matching on mel spectrogram," *Expert Systems with Applications*, vol. 213, p. 118963, 2023.

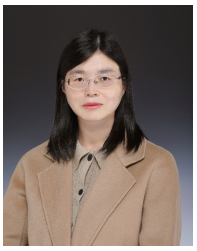
[86] Z. Xie, W. Lu, X. Liu, Y. Xue, and Y. Yeung, "Copy-move detection of digital audio based on multi-feature decision," *Journal of information security and applications*, vol. 43, pp. 37–46, 2018.

[87] Z. Su, M. Li, G. Zhang, Q. Wu, and Y. Wang, "Robust audio copy-move forgery detection on short forged slices using sliding window," *Journal of Information Security and Applications*, vol. 75, p. 103507, 2023.

[88] F. Wang, C. Li, and L. Tian, "An algorithm of detecting audio copy-move forgery based on dct and svd," in *2017 IEEE 17th International Conference on Communication Technology (ICCT)*. IEEE, 2017, pp. 1652–1657.

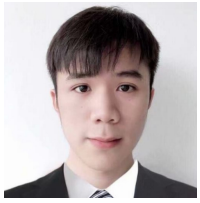
[89] W. Zhao, Y. Zhang, Y. Wang, and S. Zhang, "An audio copy-move forgery localization model by cnn-based spectral analysis," *Applied Sciences*, vol. 14, no. 11, p. 4882, 2024.

Jiayi He received the Ph.D. degree from Beijing Jiaotong University, Beijing, China, in 2021, and the B.S. degree from Wuhan University of Technology, Wuhan, China, in 2016. From 2019 to 2020, she was a visiting student at Harvard University, Boston, MA, USA. She is currently an Assistant Researcher with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China. Her research interests include fake audio detection and audio forensics.



Jiangyan Yi (Member, IEEE) received the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2018, and the M.A. degree from the Graduate School of Chinese Academy of Social Sciences, Beijing, China, in 2010. From 2011 to 2014, she was a Senior R&D Engineer with Alibaba Group. From 2018 to 2024, she was an Associate Professor with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China. She is currently an Associate Researcher with the Department of Automation, Tsinghua University. Her research interests include speech signal processing, speech recognition and synthesis, fake audio detection, audio forensics, and transfer learning.

Jianhua Tao (Senior Member, IEEE) received the M.S. degree from Nanjing University, Nanjing, China, in 1996, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2001. He is currently a Professor with Department of Automation, Tsinghua University, Beijing, China. He has authored or coauthored more than 300 papers on major journals and proceedings including the IEEE TASLP, IEEE TAFFC, IEEE TIP, IEEE TSMCB, Information Fusion, etc. His current research interests include speech recognition and synthesis, affective computing, and pattern recognition. He is the Board Member of ISCA, the chairperson of ISCA SIG-CSLP, the Chair or Program Committee Member for several major conferences, including Interspeech, ICPR, ACII, ICMI, ISCSLP, etc. He was the subject editor for the Speech Communication, and is an Associate Editor for Journal on Multimodal User Interface and International Journal on Synthetic Emotions. He was the recipient of several awards from important conferences, including Interspeech, NCMMSC, etc.



Siding Zeng received the B.E. degree from Sichuan Agricultural University, Chengdu, China, in 2021. He is currently a Master's candidate jointly supervised by the University of Chinese Academy of Sciences and State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include audio deepfake detection, unsupervised domain adaptation, and multimodal learning.

Hao Gu received the B.S. degree from Harbin Institute of Technology, Harbin, China, in 2022. He is currently a Ph.D. candidate at State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interest include fake audio detection.