

WWW: Where, Which and Whatever Enhancing Interpretability in Multimodal Deepfake Detection

Juho Jung^{1*}, Sangyoun Lee^{2*}, Jooeon Kang^{2*} and Yunjin Na^{3*}

¹Sungkyunkwan University, ²Sogang University, ³Seoul National University
jhjeon9@g.skku.edu, leesy0882@sogang.ac.kr, jekang@sogang.ac.kr, lumierej@snu.ac.kr

Abstract

All current benchmarks for multimodal deepfake detection manipulate entire frames using various generation techniques, resulting in over-saturated detection accuracies exceeding 94% at the video-level classification. However, these benchmarks struggle to detect dynamic deepfake attacks with challenging frame-by-frame alterations presented in real-world scenarios. To address this limitation, we introduce **FakeMix**, a novel clip-level evaluation benchmark aimed at identifying manipulated segments within both video and audio, providing insight into the origins of deepfakes. Furthermore, we propose novel evaluation metrics, **Temporal Accuracy (TA)** and **Frame-wise Discrimination Metric (FDM)**, to assess the robustness of deepfake detection models. Evaluating state-of-the-art models against diverse deepfake benchmarks, particularly **FakeMix**, demonstrates the effectiveness of our approach comprehensively. Specifically, while achieving an Average Precision (AP) of 94.2% at the video-level, the evaluation of the existing models at the clip-level using the proposed metrics, TA and FDM, yielded sharp declines in accuracy to 53.1%, and 52.1%, respectively. Code is available at <https://github.com/lsy0882/FakeMix>.

1 Introduction

Rapid advances in hyper-realistic deepfake technology [Zhang, 2022; Seow *et al.*, 2022; Guarnera *et al.*, 2020; Singh *et al.*, 2020; Korshunova *et al.*, 2017] have raised significant privacy and social concerns [Chen *et al.*, 2022; Li *et al.*, 2021a], requiring robust detection methods across both video [Tolosana *et al.*, 2020] and audio domains [Jia *et al.*, 2018]. Despite progress in multimodal deepfake detection, existing benchmarks such as FakeAVCeleb [Khalid *et al.*, 2021b], DFDC [Dolhansky *et al.*, 2020] and KoDF [Kwon *et al.*, 2021], focus on full-video manipulations, which often results in inflated detection accuracy and a lack of insight into specific manipulated segments. This gap

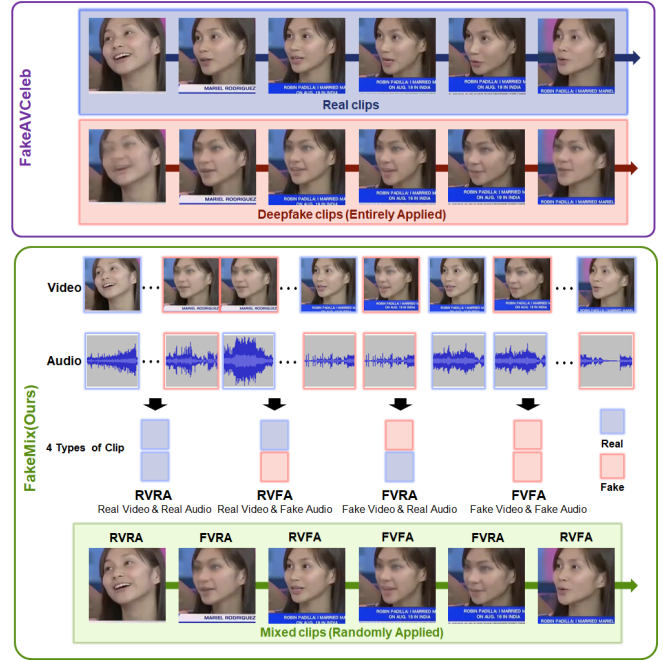


Figure 1: Comparison between previous benchmark and the proposed benchmark, **FakeMix**. While **FakeAVCeleb** operated deepfake on complete video or audio segments for video-level classification, **FakeMix** introduces dynamic frame-level alterations to enhance evaluation of deepfake video detection.

highlights the need for more precise detection methodologies that can identify and analyze the manipulated regions within the media.

Recent works have evolved from image-based methods [Khan and Dai, 2021; Tarasiou and Zafeiriou, 2020; Zheng *et al.*, 2021; Zhang *et al.*, 2021; Gu *et al.*, 2022b; Gu *et al.*, 2022a; Heo *et al.*, 2023], which analyze facial information [Ikram *et al.*, 2023; Heo *et al.*, 2023] or morphological details [Tarasiou and Zafeiriou, 2020; Li *et al.*, 2021b], to more sophisticated video-based methods [Yang *et al.*, 2023; Wang *et al.*, 2022; Lewis *et al.*, 2020; Khalid *et al.*, 2021b; Shahzad *et al.*, 2022; Hashmi *et al.*, 2022; Cai *et al.*, 2022; Khalid *et al.*, 2021a] that incorporate temporal data [Gu *et al.*, 2022a]. However, these approaches generally rely on binary classification of the entire videos (video-level), over-

*These authors contributed equally to this work.

Table 1: Comparison of Benchmark Datasets for Deepfake Detection

Dataset	Fake Video	Fake Audio	Fine-grained labeling	Deepfake Appliance
DFDC [Dolhansky <i>et al.</i> , 2020]	Yes	Yes	No	Entirely applied
KoDF [Kwon <i>et al.</i> , 2021]	Yes	No	No	Entirely applied
FakeAVCeleb [Khalid <i>et al.</i> , 2021b]	Yes	Yes	Yes	Entirely applied
FakeMix (Ours)	Yes	Yes	Yes	Randomly applied to specific segments

looking specific regions of manipulation and thereby limiting a comprehensive assessment of detection model performance. Moreover, as deepfakes become more sophisticated, the importance of temporal information in identifying inconsistencies in facial movements [Jung *et al.*, 2023] has become more pronounced, highlighting the shortcomings of current methodologies. Recognizing these challenges, our work introduces three main contributions as follows:

1. We propose *FakeMix*, a novel clip-level audio-video multimodal deepfake detection benchmark. Unlike established benchmarks that dominantly focus on overall video-level manipulation, *Fakemix* provides a distinctive assessment by pinpointing specific tampered segments within contents. This approach addresses critical limitations of current benchmarks, which overlook localized alterations.
2. We develop novel evaluation metrics, namely Temporal Accuracy (*TA*) and Frame-wise Discrimination Metric (*FDM*), designed to validate the robustness of deepfake detection models. These metrics enable precise identification of deepfake-affected regions, enhancing the granularity of results. Our comprehensive evaluation against existing benchmarks demonstrates the efficacy and necessity of incorporating *TA* and *FDM* into the evaluation framework.
3. To the best of our knowledge, this is the first attempt to assess deepfake video detection at the clip-level, aiming to enhance interpretability. By precisely identifying the specific location (**Where**), modality (**Which**), and deepfake generation technique (in **Whatever** benchmarks) employed in the manipulation, our approach represents a significant rectification, offering insights into understanding multimodal deepfake detection.

2 Related Work

There have been numerous research works that studied how to detect deepfakes in multimedia. Recently, deepfake detection studies leverage various DNN architectures to identify and distinguish manipulated videos [Nguyen *et al.*, 2022; Yu *et al.*, 2021]. Depending on which modalities are involved, deepfake detection tasks can be divided as follows:

Single-Modality Deepfake Video Detection. In general, conventional methods utilized a single modality, especially visual domain. [Li *et al.*, 2020] addressed the challenge of partial face manipulations, where only video-level labels are provided. [Gu *et al.*, 2021] exploited spatial-temporal inconsistency appeared in forged videos. To tackle the poor generalization issue, [Cozzolino *et al.*, 2021] enhanced robustness through metric learning with adversarial training to cap-

ture temporal facial features, which incorporates high-level semantic features. In spite of their effectiveness, they do not guarantee the high performance of videos with audio deepfakes all at once. This demonstrates the need for a methodology that utilizes both modalities simultaneously.

Audio-visual Deepfake Video Detection. The emergence of multimodal learning has led to the development of deepfake detection works integrating both auditory and visual modalities. [Zhou and Lim, 2021] presented a task for joint audio-video deepfake detection, leveraging intrinsic synchronization between modalities. They improved generalization abilities in unseen deepfake types, focusing on modality relationships. To this end, [Zhao *et al.*, 2022] introduced a self-supervised transformer-based contrastive learning. They leveraged learning lip motion without extensive annotations, encouraging alignment of paired audio-visual representations while promoting diversity on unpaired instances. [Feng *et al.*, 2023] developed an auto-regressive model to generate audio-visual feature sequences, capturing temporal synchronization. [Yu *et al.*, 2023] introduced a unified modality-agnostic approach to handle missing modality scenarios and extract speech correlation, making deepfakes challenging to reproduce. [Raza and Malik, 2023] also proposed a unified framework, which extracts and fuses learned channels from audio and video for effective multi-label detection. While these studies have exploited significant techniques to improve detection performance, they can only detect whether deepfakes are occurred in the entire video or audio unit.

Existing Benchmarks of Deepfake Detection. The evolution of deepfake detection has been highlighted by key benchmarks, summarized in Table 1, including DFDC [Dolhansky *et al.*, 2020], KoDF [Kwon *et al.*, 2021], and FakeAVCeleb [Khalid *et al.*, 2021b]. However, in response to the increasing complexity of deepfake techniques, they are showing obvious limitations. Existing benchmarks only involve scenarios where deepfakes are applied to every frame within the video, which cannot fully represent various real-world applications, where deepfakes can be applied to specific segments of the video. They also lack attention to delicate manipulations, such as minor changes in facial expressions or specific features. Although KoDF and FakeAVCeleb have attempted to incorporate culturally specific representations and audio-visual elements, the problem of detecting partial deepfakes remains unresolved.

3 Methodology

3.1 FakeMix

To mitigate constraints in existing benchmarks, our work introduces a new benchmark *FakeMix*, a novel clip-level assessment technique for evaluating the robustness and generaliza-

Table 2: Comparison of the performance of deepfake detection models on the established deepfake benchmark, FakeAVCeleb and the proposed benchmark, *FakeMix*.

Benchmark	Model	Modality	Task	Acc	TA	FDM
FakeAVCeleb	Xception [Khalid <i>et al.</i> , 2021b]	A	video-level	0.7306	-	-
FakeMix	Xception [Khalid <i>et al.</i> , 2021b]	A	clip-level	-	0.5905	0.6018
FakeAVCeleb	Xception [Khalid <i>et al.</i> , 2021b]	V	video-level	0.7626	-	-
FakeMix	Xception [Khalid <i>et al.</i> , 2021b]	V	clip-level	-	0.5060	0.5034
FakeAVCeleb	AVAD [Feng <i>et al.</i> , 2023]	A-V	video-level	0.9420	-	-
FakeMix	AVAD [Feng <i>et al.</i> , 2023]	A-V	clip-level	-	0.5312	0.5212

tion of multimodal deepfake detection. *FakeMix* is designed to address sophisticated scenarios where deepfakes are randomly applied to specific segments of the video and audio, offering more realistic conditions and emphasizing multimodal alignment to enhance interpretability of deepfake detection models. Unlike previous benchmarks, as depicted in Figure 1, *FakeMix* incorporates random segment insertions in clips by manipulating both video and audio within one-second intervals to measure the adaptability of multimodal deepfake detection models.

3.2 Generation and Description of FakeMix

As shown in Figure 1, let $V_r = \{v_{r1}, v_{r2}, \dots, v_{rn}\}$ and $V_f = \{v_{f1}, v_{f2}, \dots, v_{fm}\}$ represent the sets of clips from Real Video and Fake Video, respectively, where v_{ri} and v_{fi} denote the i -th clip in each set. Similarly, let $A_r = \{a_{r1}, a_{r2}, \dots, a_{rn}\}$ and $A_f = \{a_{f1}, a_{f2}, \dots, a_{fm}\}$ represent the sets of clips from Real Audio and Fake Audio, respectively, where a_{ri} and a_{fi} represent the i -th clip in each set. To create a *FakeMix* video sequence V , we randomly select clips from either V_r or V_f and concatenate them. Similarly, to create a *FakeMix* audio sequence A , we randomly select clips from either A_r or A_f and concatenate them as follows:

- Randomly selecting clips for the video sequence:

$$V = \{v_{ij} \mid v_{ij} \in V_r \cup V_f\} \quad (1)$$

- Randomly selecting clips for the audio sequence:

$$A = \{a_{ij} \mid a_{ij} \in A_r \cup A_f\} \quad (2)$$

Here, i and j denote the indices of the selected clips from the respective sets.

Consequently, within *FakeMix*, the videos are categorized at the clip-level as Real or Fake for both video and audio. Hence, the generated videos by *FakeMix* can be utilized to determine the segments within a video where deepfake manipulation occurs in both video and audio components.

3.3 Evaluation Metrics for FakeMix

As illustrated in Figure 2, we employ two novel metrics designed to offer a more granular analysis of deepfake detection capabilities: *Temporal Accuracy (TA)* and *Frame-wise Discrimination Metric (FDM)*. These metrics allow us to assess the effectiveness of deepfake detection at the individual frame level, which is critical for identifying and comprehending the temporal dynamics of deepfake manipulations.

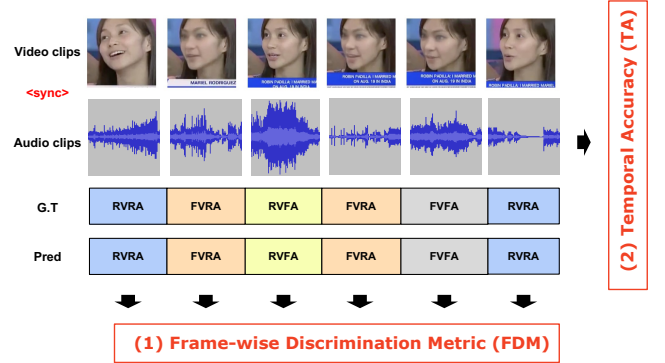


Figure 2: Comprehensive overview of *Temporal Accuracy* and *Frame-wise Discrimination Metric* conducted on the *Fakemix*.

Temporal Accuracy (TA)

TA is a metric utilized to gauge the frame-level precision of deepfake detection models in predicting the authenticity of each frame within a video. This metric is defined as:

$$TA = \frac{1}{N} \sum_{i=1}^N \left(\frac{1}{F_{v_i}} \sum_{j=1}^{F_{v_i}} I(\hat{y}_{ij} = y_{ij}) \right), \quad (3)$$

where N is the number of videos, F_{v_i} is the total number of frames in the i -th video, \hat{y}_{ij} is the predicted label for the j -th frame of the i -th video, y_{ij} is the ground truth, and I is the indicator function which is 1 if the predicted label equals the ground truth and 0 otherwise.

Frame-wise Discrimination Metric (FDM)

To complement TA, we introduce the FDM, which assesses the model's discrimination accuracy over the entire dataset at the frame level. It is expressed as:

$$FDM = \frac{\sum_{i=1}^N \sum_{j=1}^{F_{v_i}} I(\hat{y}_{ij} = y_{ij})}{\sum_{i=1}^N F_{v_i}}, \quad (4)$$

In this equation, N represents the number of videos, and F_{v_i} signifies the count of frames within the i -th video. Here, $I(\hat{y}_{ij} = y_{ij})$ computes the correctness of predictions at the individual frame level.

These metrics are pivotal as they provide a detailed understanding of a model's ability to discern real and fake content at a granular level, echoing the need for sophisticated evaluation in the age of advanced deepfakes. With TA and FDM, we

aim to establish a standard that can effectively measure and guide the development of next-generation deepfake detection models.

4 Experiments

The difference in evaluation results between FakeAVCeleb, which evaluates the models at the video level, and FakeMix, which evaluates at the clip level, demonstrates that FakeMix is more suitable for assessing the robustness and generalization of deepfake detection models.

4.1 Experimental Settings

As shown in Table 2, we first conducted an experiment to identify the differences in evaluation methodologies within a single modality by assessing the same model across each dataset. During this experiment, we utilized the Xception model previously used by [Khalid *et al.*, 2021b], and maintained consistent data preprocessing and hyperparameter settings.

Subsequently, we evaluated the robustness of the AVAD model, as proposed by [Feng *et al.*, 2023], across different modalities by testing it on each dataset. Data preprocessing for the FakeAVCeleb dataset, which contains longer video sequences, involved using sequences of length $N = 50$ from 2.0-second videos. In contrast, for the shorter 1-second clips of the FakeMix dataset, we adjusted the sequence length to $N = 50$ from 1.0-second videos. To address the reduced video duration, we scaled the probability scores output by the AVAD model, considering scores of 0.5 or higher as indicative of fakes. This scaling was critical for accurate computation of True Acceptance (TA) and False Detection Metrics (FDM). Aside from these dataset-specific preprocessing modifications, all other experimental settings conformed to those outlined by [Feng *et al.*, 2023].

4.2 Results

In the FakeAVCeleb benchmark, which tests video-level classification, the Xception model achieves approximately 76% accuracy. Meanwhile, another model, AVAD, reaches a higher accuracy of 94% in the same video-level classification on FakeAVCeleb. However, in the FakeMix benchmark, which tests clip-level classification, both Xception and AVAD models show a reduction in accuracy to around 50-60%. Notably, while the Xception model demonstrated lower performance than AVAD at the video level, it outperforms AVAD in clip-level performance.

The experimental results indicate that video-level classification often leads to overestimation, as the entire video is labeled as deepfake even if only a portion of the video contains manipulated content. Therefore, evaluating deepfake detection models at the clip or frame level is crucial to accurately verify their effectiveness. Our proposed evaluation metrics offer a more significant understanding and interpretability in deepfake detection, enabling precise identification of manipulated segments within contents. This approach enhances the reliability and applicability of deepfake detection models in practice.

5 Conclusion

The surge in hyper-realistic deepfake techniques has raised concerns about the authenticity of video and audio content. However, existing multimodal deepfake benchmarks often overlook specific manipulated segments, resulting in inflated detection accuracy and a lack of insight. To address this exaggerated efficacy, we introduced *FakeMix*, a clip-level evaluation benchmark that enhances interpretability by targeting manipulated video-audio segments. Additionally, our proposed evaluation metrics, *TA* and *FDM*, effectively assess the robustness and reliability of deepfake detection methods. By rethinking the overall assessment framework, these findings which highlight the importance of adopting clip-level assessments and refined evaluation metrics, lay the groundwork for more comprehensive and accurate deepfake detection strategies to combat deceptive content.

Ethical Statement

There are no ethical issues.

Acknowledgments

This research was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by Korea government (MSIT) (RS-2022-00143911, AI Excellence Global Innovative Leader Education Program).

References

- [Cai *et al.*, 2022] Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar 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)*, pages 1–10. IEEE, 2022.
- [Chen *et al.*, 2022] Zhaoyu Chen, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Wenqiang Zhang. Towards practical certifiable patch defense with vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15148–15158, 2022.
- [Cuzzolino *et al.*, 2021] Davide Cuzzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15108–15117, 2021.
- [Dolhansky *et al.*, 2020] Brian Dolhansky, Joanna Bitton, Ben Pfau, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.
- [Feng *et al.*, 2023] Chao Feng, Ziyang Chen, and Andrew Owens. Self-supervised video forensics by audio-visual anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10491–10503, 2023.

- [Gu *et al.*, 2021] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3473–3481, 2021.
- [Gu *et al.*, 2022a] Zhihao Gu, Taiping Yao, Yang Chen, Shouhong Ding, and Lizhuang Ma. Hierarchical contrastive inconsistency learning for deepfake video detection. In *European Conference on Computer Vision*, pages 596–613. Springer, 2022.
- [Gu *et al.*, 2022b] Zhihao Gu, Taiping Yao, C Yang, Ran Yi, Shouhong Ding, and Lizhuang Ma. Region-aware temporal inconsistency learning for deepfake video detection. In *Proceedings of the 31th International Joint Conference on Artificial Intelligence*, volume 1, 2022.
- [Guarnera *et al.*, 2020] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 666–667, 2020.
- [Hashmi *et al.*, 2022] Ammarah Hashmi, Sahibzada Adil Shahzad, Wasim Ahmad, Chia Wen Lin, Yu Tsao, and Hsin-Min Wang. Multimodal forgery detection using ensemble learning. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1524–1532. IEEE, 2022.
- [Heo *et al.*, 2023] Young-Jin Heo, Woon-Ha Yeo, and Byung-Gyu Kim. Deepfake detection algorithm based on improved vision transformer. *Applied Intelligence*, 53(7):7512–7527, 2023.
- [Ikram *et al.*, 2023] Sumaiya Thaseen Ikram, Shourya Chambial, Dhruv Sood, et al. A performance enhancement of deepfake video detection through the use of a hybrid cnn deep learning model. *International journal of electrical and computer engineering systems*, 14(2):169–178, 2023.
- [Jia *et al.*, 2018] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, 31, 2018.
- [Jung *et al.*, 2023] Juho Jung, Chaewon Kang, Jeewoo Yoon, Simon S Woo, and Jinyoung Han. Safe: Sequential attentive face embedding with contrastive learning for deepfake video detection. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3993–3997, 2023.
- [Khalid *et al.*, 2021a] Hasam Khalid, Minha Kim, Shahroz Tariq, and Simon S Woo. Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In *Proceedings of the 1st workshop on synthetic multimedia-audiovisual deepfake generation and detection*, pages 7–15, 2021.
- [Khalid *et al.*, 2021b] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021.
- [Khan and Dai, 2021] Sohail Ahmed Khan and Hang Dai. Video transformer for deepfake detection with incremental learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1821–1828, 2021.
- [Korshunova *et al.*, 2017] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE International Conference on computer vision*, pages 3677–3685, 2017.
- [Kwon *et al.*, 2021] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. Kodf: A large-scale korean deepfake detection dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10744–10753, 2021.
- [Lewis *et al.*, 2020] John K Lewis, Imad Eddine Toubal, Helen Chen, Vishal Sandesera, Michael Lomnitz, Zigfried Hampel-Arias, Callyam Prasad, and Kannappan Palaniappan. Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning. In *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–9. IEEE, 2020.
- [Li *et al.*, 2020] Xiaodan Li, Yining Lang, Yuefeng Chen, Xiaofeng Mao, Yuan He, Shuhui Wang, Hui Xue, and Quan Lu. Sharp multiple instance learning for deepfake video detection. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1864–1872, 2020.
- [Li *et al.*, 2021a] Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, Jilin Li, and Feiyue Huang. Detecting adversarial patch attacks through global-local consistency. In *Proceedings of the 1st International Workshop on Adversarial Learning for Multimedia*, pages 35–41, 2021.
- [Li *et al.*, 2021b] Meng Li, Beibei Liu, Yongjian Hu, Liepiao Zhang, and Shiqi Wang. Deepfake detection using robust spatial and temporal features from facial landmarks. In *2021 IEEE International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2021.
- [Nguyen *et al.*, 2022] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M Nguyen. Deep learning for deep-fakes creation and detection: A survey. *Computer Vision and Image Understanding*, 223:103525, 2022.
- [Raza and Malik, 2023] Muhammad Anas Raza and Khalid Mahmood Malik. Multimodaltrace: Deepfake detection using audiovisual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 993–1000, 2023.
- [Seow *et al.*, 2022] Jia Wen Seow, Mei Kuan Lim, Raphael CW Phan, and Joseph K Liu. A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing*, 513:351–371, 2022.
- [Shahzad *et al.*, 2022] Sahibzada Adil Shahzad, Ammarah Hashmi, Sarwar Khan, Yan-Tsung Peng, Yu Tsao, and

- Hsin-Min Wang. Lip sync matters: A novel multimodal forgery detector. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1885–1892. IEEE, 2022.
- [Singh *et al.*, 2020] Simranjeet Singh, Rajneesh Sharma, and Alan F Smeaton. Using gans to synthesise minimum training data for deepfake generation. *arXiv preprint arXiv:2011.05421*, 2020.
- [Tarasiou and Zafeiriou, 2020] Michail Tarasiou and Stefanos Zafeiriou. Extracting deep local features to detect manipulated images of human faces. In *2020 IEEE international conference on image processing (ICIP)*, pages 1821–1825. IEEE, 2020.
- [Tolosana *et al.*, 2020] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020.
- [Wang *et al.*, 2022] Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Yu-Gang Jiang, and Ser-Nam Li. M2tr: Multi-modal multi-scale transformers for deepfake detection. In *Proceedings of the 2022 international conference on multimedia retrieval*, pages 615–623, 2022.
- [Yang *et al.*, 2023] Wenyuan Yang, Xiaoyu Zhou, Zhikai Chen, Bofei Guo, Zhongjie Ba, Zhihua Xia, Xiaochun Cao, and Kui Ren. Avoid-df: Audio-visual joint learning for detecting deepfake. *IEEE Transactions on Information Forensics and Security*, 18:2015–2029, 2023.
- [Yu *et al.*, 2021] Peipeng Yu, Zhihua Xia, Jianwei Fei, and Yujiang Lu. A survey on deepfake video detection. *Iet Biometrics*, 10(6):607–624, 2021.
- [Yu *et al.*, 2023] Cai Yu, Peng Chen, Jiahe Tian, Jin Liu, Jiao Dai, Xi Wang, Yesheng Chai, and Jizhong Han. Modality-agnostic audio-visual deepfake detection. *arXiv preprint arXiv:2307.14491*, 2023.
- [Zhang *et al.*, 2021] Daichi Zhang, Chenyu Li, Fanzhao Lin, Dan Zeng, and Shiming Ge. Detecting deepfake videos with temporal dropout 3dcnn. In *IJCAI*, pages 1288–1294, 2021.
- [Zhang, 2022] Tao Zhang. Deepfake generation and detection, a survey. *Multimedia Tools and Applications*, 81(5):6259–6276, 2022.
- [Zhao *et al.*, 2022] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Weiming Zhang, and Nenghai Yu. Self-supervised transformer for deepfake detection. *arXiv preprint arXiv:2203.01265*, 2022.
- [Zheng *et al.*, 2021] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15044–15054, 2021.
- [Zhou and Lim, 2021] Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14800–14809, 2021.