Interspeech 2024
1-5 September 2024, Kos, Greece

Zero-Shot Fake Video Detection by Audio-Visual Consistency

Xiaolou Li[1], *Zehua Liu*[1], *Chen Chen*[2], *Lantian Li*[1], *Li Guo*[1], *Dong Wang*[2]

[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China
[2]Center for Speech and Language Technologies, BNRist, Tsinghua University, China

Corresponding authors: lilt@bupt.edu.cn, wangdong99@mails.tsinghua.edu.cn

Abstract

Recent studies have advocated the detection of fake videos as a one-class detection task, predicated on the hypothesis that the consistency between audio and visual modalities of genuine data is more significant than that of fake data. This methodology, which solely relies on genuine audio-visual data while negating the need for forged counterparts, is thus delineated as a 'zero-shot' detection paradigm. This paper introduces a novel zero-shot detection approach anchored in content consistency across audio and video. By employing pre-trained ASR and VSR models, we recognize the audio and video content sequences, respectively. Then, the edit distance between the two sequences is computed to assess whether the claimed video is genuine. Experimental results indicate that, compared to two mainstream approaches based on semantic consistency and temporal consistency, our approach achieves superior generalizability across various deepfake techniques and demonstrates strong robustness against audio-visual perturbations. Finally, state-of-the-art performance gains can be achieved by simply integrating the decision scores of these three systems.

Index Terms: fake video detection, zero-shot, audio-visual

1. Introduction

In recent years, the development of deepfake technologies has made it possible to generate high-fidelity fake videos [1, 2, 3, 4]. These technologies leverage advanced methods such as face-swapping, lip-syncing for video generation, and speech synthesis or voice conversion for audio generation. These fake videos pose significant risks of misleading the public, damaging reputations, threatening security, and undermining trust [5, 6, 7]. Consequently, developing deepfake detection technologies has emerged as a critical concern.

As deepfake advances, the development of countermeasures has concurrently evolved [8, 9, 10, 11, 12]. Initially, deepfakes primarily relied on face-swapping techniques, producing fake videos characterized by unnaturally smooth areas within frames or notable discontinuities between frames. Consequently, early fake video detection strategies identified these forgeries by recognizing such artifacts. For example, Zheng et al. [13] proposed detecting forgeries by capturing the discontinuities between video frames. Haliassos et al. [14] achieved deepfake detection by identifying semantic irregularities in the lip movements within videos.

However, with the further advancement of deepfake technologies, relying solely on the video modality for fake video detection has become exceedingly challenging. To address this challenge, researchers expanded their focus to introduce audio modality to assist in fake video detection, leading to audio-visual multi-modal forgery detection. Initially, researchers adopted an end-to-end binary classification framework to discriminate between genuine and fake videos. For instance, Wang et al. [15] proposed a multi-modal detection network that takes raw audio and video streams as input. By leveraging an attention mechanism, this network integrates audio and video features deeply, ultimately distinguishing between genuine and fake videos using a binary classifier.

Although these approaches demonstrated preliminary effectiveness, their fundamental limitation was the independent detection of artifacts in audio and video without considering the impact of deepfakes on the consistency between audio and video. Genuine videos naturally possess intrinsic consistency between audio and video modalities, while deepfakes may somewhat corrupt this consistency. Therefore, several studies have focused on evaluating the consistency between audio and video for deepfake detection. For example, Shahzad et al. [16] detected forgeries by quantifying the mismatch between the lip sequence extracted from the video and the synthetic lip sequence generated from the audio by the Wav2Lip model [17]. Chugh et al. [18] introduced a contrastive loss to enforce similarity in audio and video representations of genuine video pairs and dissimilarity in those of fake pairs, thereby establishing inter-modality similarity. Zhang et al. [19] adopted the same strategy. Cheng et al. [20] argued that there is a high homogeneity between a person's face and voice. They, therefore, detect fake videos by assessing the matching degree between face and voice representations.

Despite the effectiveness of incorporating audio-visual consistency in improving detection performance, these methods generally rely on an end-to-end two-class classification framework. This framework performs well for detecting *specific* deepfakes but lacks generalizability for unseen deepfakes. Considering that the consistency of audio-visual modalities is an intrinsic property of genuine videos, this two-class classification task can be re-conceptualized as a one-class detection task that solely detects whether a video is genuine. Notably, this one-class framework requires only genuine audio-visual data for modelling, without the need for any fake data, thus regarded as 'zero-shot' deepfake detection. For instance, Cozzolino et al. [21] utilized face recognition and speaker recognition models trained on genuine data to detect forgeries by assessing the consistency between face identity and speaker identity. Feng et al. [22] trained solely on genuine audio-visual data and detected forgeries by assessing the temporal synchrony between audio and video. Tal et al. [23] used the AV-HuBERT model [24] trained on real audio-visual data to detect forgeries by quan-

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants No.62301075/62171250. X. Li and Z. Liu are joint first authors.

10.21437/Interspeech.2024-2497

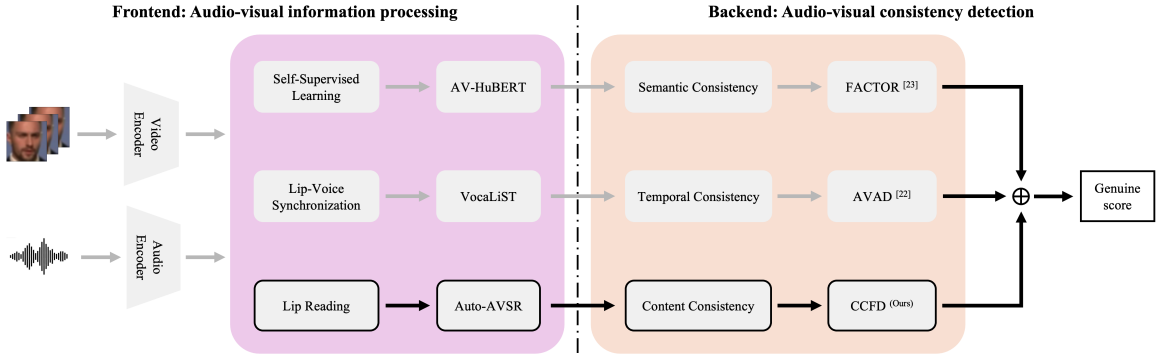

Figure 1: *The unified framework of zero-shot fake video detection.*

tifying the distance between semantic representations of audio and video. Since these zero-shot approaches only require genuine data for modelling, they can detect *any* type of deepfakes, demonstrating stronger generalization capabilities.

This paper introduces a novel audio-visual deepfake detection method based on voice-lip content consistency. The fundamental assumption of our method is that genuine data possess intrinsic content consistency between voice streams and lip movements. To this end, we first employ automatic speech recognition (ASR) and visual speech recognition (VSR) models trained on genuine data to decode the content sequences from audio and video, respectively. We then compute the edit distance between the two content sequences as a metric to evaluate the degree of content consistency between the modalities. Experimental results demonstrate that compared to semantic-consistency-based FACTOR [23] and temporal-consistency-based AVAD [22], our content-consistency method achieves great performance across a variety of deepfake datasets (including FakeAVCeleb [25] and DeepFakeTIMIT [26]), showcasing superior generalizability. Moreover, we ascertain that our method is more robust by introducing various perturbations into both audio and video. Finally, believing in the complementary strengths of the three distinct consistency approaches, we advocate for a simple score fusion approach to combine these methods. Our results indicate that this fusion achieves state-of-the-art (SOTA) performance in deepfake detection, setting a new benchmark in the field.

2. Zero-Shot Fake Video Detection

In this section, we delineate a unified framework for zero-shot fake video detection, as shown in Figure 1. This framework is composed of two components: frontend audio-visual information processing and backend audio-visual consistency detection.

2.1. Frontend: Audio-visual information processing

At the frontend, various audio-visual information processing tasks employ different model architectures, training paradigms, and objectives. Despite these differences, the underlying methodology remains consistent, involving encoding genuine audio and visual inputs to derive their respective latent representations, then establishing the correlation between the representations of the two modalities, and finally leveraging the correlation to learn the pretext task. For instance, AV-HuBERT [24] adopts a self-supervised learning strategy, engaging in an iterative process of feature clustering and learning new features via a masked prediction loss. This process uncovers strong correlations between audio streams and lip movements, yielding a highly effective pre-trained model that has been successfully deployed in various audio-visual downstream applications. Besides, VocaLiST [27] designs a powerful cross-modal Transformer model for learning the correlation across audio and visual streams. Then, it outputs a score indicating whether the voice and lip motion are synchronised. Moreover, AV-ASR [28, 29], by leveraging the correlation of contextual information across audio and visual streams, integrates audio and visual contextual representations to enhance visual speech recognition, also known as lip reading.

2.2. Backend: Audio-visual consistency detection

Considering that these audio-visual information processing frontends require only genuine audio-visual data during the training phase without the need for any fake data, these well-trained frontend models have proficiently learned the intrinsic inter-modality correlations within genuine data. A natural intuition is that these correlations observed between audio and video modalities in genuine data are much weaker in fake data. Hence, at the backend, quantifying these correlations allows for the measurement of consistency across audio and video elements, facilitating the determination of a video's authenticity. For instance, FACTOR [23] leverages a pre-trained AV-HuBERT model to extract latent representations of audio and video; it then employs cosine similarity to assess the semantic consistency between these representations, yielding a decision score. This approach represents a semantic-consistency-based method of fake video detection and has achieved great performance on the FakeAVCeleb dataset. In addition, AVAD [22] trained a model on lip-voice synchronization, generating features that describe the temporal synchronization between audio and visual streams, and subsequently predicts a consistency score based on these features, thus representing a temporal-consistency-based approach to fake video detection.

In this paper, we introduce a novel fake video detection approach based on content consistency, termed **c**ontent **c**onsistency **f**ake **d**etection, CCFD. Specifically, we posit that for genuine data, there is a strong correlation between the content information of audio and visual streams. Following this hypothesis, we leverage ASR and VSR models within the AV-ASR framework to decode the content sequences of audio streams and lip movements, respectively. The consistency between audio and video is then measured by computing the edit distance between these two content sequences. In this study, we use the audio content sequence decoded by ASR as the reference and the lip content sequence decoded by VSR as the hypothesis when computing the word error rate (WER), thereby determining if the claimed video is genuine or not.

Finally, we believe various consistency-based detection methods, grounded in different tasks and assumptions, possess inherent complementary qualities. Therefore, fusing the decision scores output by different detection methods is feasible to arrive at an improved detection assessment.

3. Experiment Settings

3.1. Data

Our experiments used two datasets: FakeAVCeleb [25] and DeepFakeTIMIT [26].

FakeAVCeleb, a large-scale audio-visual deepfake dataset. The genuine videos were selected from VoxCeleb2 [30]. It employed face-swapping algorithms such as Faceswap [31] and FSGAN [32] to generate swapped fake videos. Besides, it used an SV2TTS tool [33] to generate cloned audios. After generating fake videos and audios, Wav2Lip [17] was applied to fake videos to reenact the videos based on fake audios. In our experiments, we sampled 50 genuine videos and 2,085 fake videos from 50 celebrities for performance evaluation.

DeepFakeTIMIT, a standard deepfake dataset. It encompasses 320 genuine videos selected from VidTIMIT[1], featuring 16 pairs of speakers with similar visual characteristics. Furthermore, 320 fake videos were produced using advanced face-swapping techniques.

In our experiments, all videos were transcoded to a frame rate of 25 frames per second, and all audios were resampled to a sampling rate of 16kHz.

3.2. Systems

3.2.1. SCFD: Semantic-consistency fake detection

Followed by FACTOR[2], we constructed a semantic-consistency fake detection (SCFD) system.

Firstly, we followed the preprocessing procedure outlined by Auto-AVSR[3], which includes: (1) Utilizing RetinaFace [34] for facial landmark detection. (2) Affine transformation is applied to align and stabilize the facial region in the original video, reducing the impact of head movements and centring the mouth area. (3) After alignment and stabilization, a 96x96 pixel region centred around the mouth is cropped from the frames.

Subsequently, semantic representations for each video frame and its corresponding audio segment are independently extracted using the video and audio encoders provided by AV-HuBERT. The cosine similarity between the two semantic representations is computed, resulting in a semantic consistency score for each frame.

Finally, we take the 3rd percentile of scores from all frames as the result to obtain a video-level semantic consistency score.

3.2.2. TCFD: Temporal-consistency fake detection

Inspired by AVAD [22], we developed a temporal-consistency fake detection (TCFD) system.

Initially, we adhere to the preprocessing protocol established by VocaLiST[4], utilizing facial detection techniques to extract the facial region from the original videos. After resizing the extracted region to 96x96 pixels, we specifically crop the area encompassing the lips.

Subsequently, employing a window length of five frames with a stride of one frame, the lip stream and audio stream are concurrently input into the VocaLiST pre-trained model. This procedure calculates a synchronization score for each window, reflecting the temporal alignment between the audio and video streams within that specific five-frame window.

Ultimately, to determine the video's overall temporal consistency, the average synchronization score across all windows is computed, providing a comprehensive video-level temporal consistency score.

3.2.3. CCFD: Content-consistency fake detection

In this study, we have introduced a content-consistency fake detection (CCFD) system to detect fake videos by assessing the content consistency between audio and visual streams.

The pipeline of data preprocessing is consistent with that of SCFD. We utilize Visual Speech Recognition (VSR) and Automatic Speech Recognition (ASR) models [35, 36, 37] that have been released in the Auto-AVSR repository[5]. The decoding process [38] is conducted using the BeamSearch algorithm, setting the beam size to 40.

Subsequently, video and audio streams are separately fed into the VSR and ASR models to decode the respective content sequences. Considering the superior accuracy of ASR over VSR in recognizing content, we designate the audio content sequence decoded by ASR as the reference and the lip content sequence decoded by VSR as the hypothesis. The degree of content consistency between these sequences is quantified by computing the Word Error Rate (WER), thereby offering a metric to measure the authenticity of the video.

3.2.4. System fusion

Intuitively, the three fake detection systems leverage different consistency criteria, suggesting inherent complementarity among them. Therefore, we believe that fusing the output of these systems could significantly enhance the generalizability and robustness of the final detection.

We simply average the scores from the three systems to achieve this fusion. Before this fusion, it is essential to normalize the score from each system to standardize the value range. In our experiments, the min-max normalization method is utilized for SCFD and TCFD systems. For the CCFD system, score normalization is implemented using the formula 1-min(WER, 1). This normalization process ensures that the scores across different systems are comparable, facilitating a balanced and effective fusion of their outputs.

4. Experimental Results

In this section, we evaluate the generalizability and robustness of various fake detection systems, emphasizing the different impact of audio-visual consistency criteria.

4.1. Generalization tests

For generalization tests, experiments were conducted on the FakeAVCeleb and DeepFakeTIMIT datasets. Given the variety of deepfake techniques in FakeAVCeleb, we split this dataset into several subsets based on the deepfake mode and technique for detailed analysis and reported performance on each subset. Deepfake modes were categorized into three groups: RVFA (real video with fake audio), FVRA (fake video with real audio),

[1] http://conradsanderson.id.au/vidtimit/

[2] https://github.com/talreiss/FACTOR

[3] https://github.com/mpc001/auto_avsr/tree/main/preparation

[4] https://github.com/vskadandale/vocalist

[5] https://github.com/mpc001/auto_avsr/tree/main

Table 1: *AUC results for fake detection systems across datasets.*

Dataset	FakeAVCeleb							DeepFakeTIMIT	Mean	Std.
	RVFA	FVRA			FVFA					
Method		WL	GAN	FS	WL	GAN-WL	FS-WL			
SCFD	**0.9924**	0.9481	0.7741	0.7167	0.9686	0.9687	0.9656	0.9110	0.9056	0.0961
TCFD	0.7824	0.9545	0.5262	0.4887	0.9728	0.9734	0.9685	0.8144	0.8101	0.1883
CCFD	0.6880	0.9575	**0.8403**	**0.8424**	0.9522	0.9472	0.9412	0.9315	0.8875	0.0877
Fusion	0.8624	**0.9811**	0.8144	0.7704	**0.9857**	**0.9841**	**0.9799**	**0.9786**	**0.9196**	**0.0837**

and FVFA (fake video and audio), and deepfake techniques include Wav2Lip (WL), FSGAN (GAN), and FaceSwap (FS). The Area Under the Curve (AUC) was employed as the evaluation metric. We measured the mean and standard deviation of the AUC scores on different datasets and used these quantities to evaluate the generalizability of a fake detection method. Notably, higher means and lower standard deviations indicate superior generalizability. The results are reported in Table 1.

Firstly, SCFD achieved the highest mean AUC among the three consistency-based detection systems, while CCFD exhibited the smallest standard deviation. More careful analysis revealed that SCFD was most effective in the RVFA mode; TCFD excelled against WL-based techniques; and CCFD demonstrated exceptional efficacy against GAN- and FS-based techniques within the FVRA mode, highlighting a clear bias of consistency criteria towards specific deepfake modes and techniques.

Secondly, our proposed system, CCFD, showed remarkable stability across datasets except for RVFA. This can be attributed to the high fidelity of the audio synthesized by SV2TTS, which permitted the ASR model to achieve accurate results even on fake audio.

Finally, fusing the three systems resulted in enhanced accuracy and generalizability, underscoring the complementary nature of the consistency criteria in the realm of fake detection.

4.2. Robustness tests

In practical applications, videos often encounter various noises and corruptions, such as background noise and compression artifacts, which may affect the performance of fake detection systems. Thus, evaluating the robustness of a fake detection system against a range of perturbations is of paramount importance.

We implemented three levels of video perturbations using the Kornia library[6] and FFmpeg library[7], with the specifics detailed in Table 2. For audio perturbations, four types of noise were added using the Torchaudio library[8] at three signal-to-noise ratio (SNR) levels: 12.5 dB, 2.5 dB, and -7.5 dB. The results on FakeAVCeleb are depicted in Figure 2 and summarized in Table 3.

Table 2: *Three levels of video perturbations.*

Type Parameter	Blur sigma	Noise std	Contrast factor	Compression CRF
Level 1	0.1	0.01	0.8	33
Level 2	2	0.05	1.2	40
Level 3	5	0.1	2	47

It can be observed that SCFD demonstrated considerable robustness against video perturbations but was highly sensitive to audio perturbations. This implies a significant alteration in the audio representations that the AV-HuBERT model extracts upon adding audio perturbations. In contrast, TCFD showed an inverse trend, demonstrating robustness to audio perturbations and vulnerability to video perturbations. This indicates that the audio-visual synchronization detection model lacks robustness against video perturbations. The performance of our proposed CCFD lies between SCFD and TCFD, showing stability in the face of both audio and video perturbations. After integrating the three systems, a consistent advantage in robustness was achieved across all test cases. This outcome reaffirms the complementary nature of the three consistency criteria, which can be combined to construct a stronger fake detection system.

[6]https://github.com/kornia/kornia

[7]https://ffmpeg.org/

[8]https://pytorch.org/audio/stable/index.html

Table 3: *Mean and Std. of different systems in robustness tests.*

AUC	Video		Audio	
	Mean	Std.	Mean	Std.
SCFD	0.894	0.039	0.704	0.230
TCFD	0.838	0.043	0.851	**0.012**
CCFD	0.888	0.062	0.823	0.137
Fusion	**0.927**	**0.032**	**0.879**	0.083

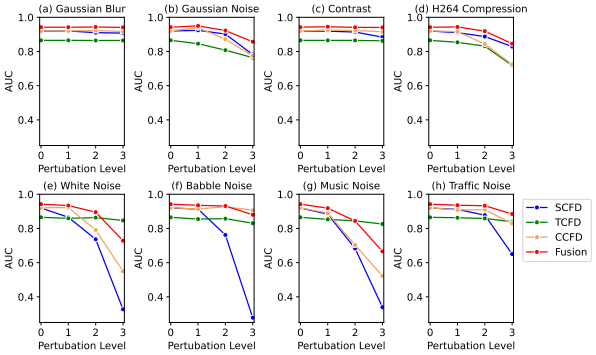

Figure 2: *AUC on FakeAVCeleb under levels of perturbations.*

5. Conclusion

This paper focuses on zero-shot fake video detection. We introduce a unified framework for zero-shot fake detection methods: an audio-visual information processing frontend and an audio-visual consistency detection backend. Following this, we constructed three zero-shot fake detection systems with different consistency criteria, including a novel method based on content consistency. Experimental results demonstrate that different systems excel at different types of deepfakes and are sensitive to different audio/video perturbations. Compared to existing methods based on temporal consistency and semantic consistency, our proposed content consistency detection system presents stable generalizability and robustness. By fusing these systems, we achieved SOTA performance on the FakeAVCeleb dataset, highlighting the complementarity among the three consistency criteria. Future work will continue exploring more consistency criteria for our zero-shot fake detection framework.

6. References

[1] M. Westerlund, "The emergence of deepfake technology: A review," *Technology Innovation Management Review*, vol. 9, no. 11, 2019.

[2] B. U. Mahmud and A. Sharmin, "Deep insights of deepfake technology: A review," *arXiv preprint arXiv:2105.00192*, 2021.

[3] A. O. Kwok and S. G. Koh, "Deepfake: a social construction of technology perspective," *Current Issues in Tourism*, vol. 24, no. 13, pp. 1798–1802, 2021.

[4] M. Sharma and M. Kaur, "A review of deepfake technology: an emerging ai threat," *Soft Computing for Security Applications: Proceedings of ICSCS 2021*, pp. 605–619, 2022.

[5] A. de Rancourt-Raymond and N. Smaili, "The unethical use of deepfakes," *Journal of Financial Crime*, vol. 30, no. 4, pp. 1066–1077, 2023.

[6] P. Singh and B. Dhiman, "Exploding ai-generated deepfakes and misinformation: A threat to global concern in the 21st century," *Authorea Preprints*, 2023.

[7] S. Choudhary, "Unmasking the truth: The rise of deepfakes and its implication on society," *Jus Corpus LJ*, vol. 4, p. 110, 2023.

[8] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE access*, vol. 10, pp. 25 494–25 513, 2022.

[9] P. Yu, Z. Xia, J. Fei, and Y. Lu, "A survey on deepfake video detection," *Iet Biometrics*, vol. 10, no. 6, pp. 607–624, 2021.

[10] 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.

[11] J. W. Seow, M. K. Lim, R. C. Phan, and J. K. Liu, "A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities," *Neurocomputing*, vol. 513, pp. 351–371, 2022.

[12] A. Deshmukh and S. B. Wankhade, "Deepfake detection approaches using deep learning: A systematic review," *Intelligent Computing and Networking: Proceedings of IC-ICN 2020*, pp. 293–302, 2020.

[13] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 15 044–15 054.

[14] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5039–5049.

[15] G. Wang, P. Zhang, L. Xie, W. Huang, Y. Zha, and Y. Zhang, "An audio-visual attention based multimodal network for fake talking face videos detection," *arXiv preprint arXiv:2203.05178*, 2022.

[16] S. A. Shahzad, A. Hashmi, S. Khan, Y.-T. Peng, Y. Tsao, and H.-M. Wang, "Lip sync matters: A novel multimodal forgery detector," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 1885–1892.

[17] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 484–492.

[18] 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 ACM International Conference on Multimedia*, 2020, pp. 439–447.

[19] Y. Zhang, W. Lin, and J. Xu, "Joint audio-visual attention with contrastive learning for more general deepfake detection," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 5, pp. 1–23, 2024.

[20] H. Cheng, Y. Guo, T. Wang, Q. Li, X. Chang, and L. Nie, "Voice-face homogeneity tells deepfake," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 3, pp. 1–22, 2023.

[21] D. Cozzolino, A. Pianese, M. Nießner, and L. Verdoliva, "Audio-visual person-of-interest deepfake detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 943–952.

[22] C. Feng, Z. Chen, and A. Owens, "Self-supervised video forensics by audio-visual anomaly detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 491–10 503.

[23] T. Reiss, B. Cavia, and Y. Hoshen, "Detecting deepfakes without seeing any," *arXiv preprint arXiv:2311.01458*, 2023.

[24] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," *arXiv preprint arXiv:2201.02184*, 2022.

[25] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "Fakeavceleb: A novel audio-video multimodal deepfake dataset," *arXiv preprint arXiv:2108.05080*, 2021.

[26] P. Korshunov and S. Marcel, "Deepfakes: a new threat to face recognition? assessment and detection," *arXiv preprint arXiv:1812.08685*, 2018.

[27] V. S. Kadandale, J. F. Montesinos, and G. Haro, "Vocalist: An audio-visual synchronisation model for lips and voices," *arXiv preprint arXiv:2204.02090*, 2022.

[28] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2021, pp. 7613–7617.

[29] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "Auto-avsr: Audio-visual speech recognition with automatic labels," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2023, pp. 1–5.

[30] J. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*. ISCA, 2018.

[31] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3677–3685.

[32] Y. Nirkin, Y. Keller, and T. Hassner, "Fsgan: Subject agnostic face swapping and reenactment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7184–7193.

[33] 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.

[34] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5203–5212.

[35] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[37] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with lstms for lipreading," *arXiv preprint arXiv:1703.04105*, 2017.

[38] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic, "Audio-visual speech recognition with a hybrid ctc/attention architecture," in *IEEE Spoken Language Technology Workshop*. IEEE, 2018, pp. 513–520.