Multimodal Deepfake Detection for Short Videos

Abderrazzaq Moufidi[1,2], David Rousseau[2] and Pejman Rasti[1,2]
[1]*Centre d'Études et de Recherche pour l'Aide à la Décision (CERADE), ESAIP, 18 Rue du 8 Mai 1945, Saint-Barthélemy-d'Anjou 49124, France*
[2]*Laboratoire Angevin de Recherche en Ingénierie des Systèmes (LARIS), UMR INRAe-IRHS, Université d'Angers, 62 Avenue Notre Dame du Lac, Angers 49000, France*

Keywords: Deepfake, Multimodality, Multi-View, Audio-Lips Correlation, Late Fusion, Spatiotemporal.

Abstract: The focus of this study is to address the growing challenge posed by AI-generated, persuasive but often misleading multimedia content, which poses difficulties for both human and machine learning interpretation. Building upon our prior research, we analyze the visual and auditory elements of multimedia to identify multimodal deepfakes, with a specific focus on the lower facial area in video clips. This targeted approach sets our research apart in the complex field of deepfake detection. Our technique is particularly effective for short video clips, lasting from 200 milliseconds to one second, surpassing many current deep learning methods that struggle in this duration. In our previous work, we utilized late fusion for correlating audio and lip movements and developed a novel method for video feature extraction that requires less computational power. This is a practical solution for real-world applications with limited computing resources. By adopting a multi-view strategy, the proposed network can leverage various weaknesses found in deepfake generation, from visual anomalies to motion inconsistencies or issues with jaw positioning, which are common in such content.

1 INTRODUCTION

In an era marked by the rapid advancements in artificial intelligence, the democratization and sophistication of multimedia content manipulation have opened up new creative possibilities and practical applications (Masood et al., 2023; Zhang, 2022). However, this increased accessibility also poses a significant challenge in detecting multimedia manipulations that are becoming increasingly elusive, with potential consequences for both human judgment and automated systems (Ilyas et al., 2023; Prajwal et al., 2020; Zhou and Lim, 2021). Specifically, the manipulation of multimedia content, such as convincing deepfakes designed to disseminate false information and fake news, involves meticulous alterations of both video and audio channels. For instance, a deepfake featuring a politician delivering a speech may employ techniques like lip sync for video modification and the use of an impersonator's voice (Prajwal et al., 2020; Ling et al., 2022). With recent advancements in text-to-speech (TTS) (Jiang et al., 2023; Jia et al., 2018) and voice conversion (VC) algorithms (Huang et al., 2021), the synthesis of human speech is becoming increasingly accessible, suggesting a future where audio plays an equally crucial role as video in deepfake detection.

Another commonly employed technique is face swapping, wherein an individual's facial features in a video undergo a seamless replacement with those of another person, often leveraging Generative Adversarial Networks (GANs) to achieve photorealistic outcomes (Sanderson and Lovell, 2009; Korshunova et al., 2017; Nirkin et al., 2019). Voice conversion serves as an extension to the capabilities of deepfakes, enabling the alteration of one's voice to closely emulate the vocal characteristics of another individual (Huang et al., 2021). When integrated with visual manipulations, this technology has the potential to generate even more compelling synthetic media.

Text-to-speech synthesis has achieved a high level of realism, making synthetic voices increasingly difficult to distinguish from human ones. This advancement allows the generation of lifelike audio content from text inputs (Jiang et al., 2023; Jia et al., 2018). While these developments are remarkable, they highlight the necessity for robust detection mechanisms to ensure the credibility of multimedia channels in the evolving landscape of synthetic media.

Conventional approaches to identify video manipulations frequently hinge on 3D deep learning models (Zhou and Lim, 2021; Zi et al., 2020), which

Moufidi, A., Rousseau, D. and Rasti, P.
Multimodal Deepfake Detection for Short Videos.
DOI: 10.5220/0012557300003720

In *Proceedings of the 4th International Conference on Image Processing and Vision Engineering (IMPROVE 2024)*, pages 67-73
ISBN: 978-989-758-693-4; ISSN: 2795-4943

may pose practical challenges in real-world situations characterized by limited computational resources. Alternatively, some methods resort to 2D models applied to video frames, overlooking the temporal information inherent in videos (Ilyas et al., 2023; Zi et al., 2020; Khalid et al., 2021a).

An overlooked and crucial aspect in this domain is the identification of manipulated short utterances within both visual video and audio content (Zhou and Lim, 2021). The limited data available in short sequences amplifies the difficulty of reliable detection, as algorithms face challenges in discerning patterns from constrained information. This underscores the pressing requirement for models that are both efficient and effective. Notably, there is a notable absence of existing efforts specifically dedicated to addressing the challenge of detecting deepfakes in short utterance audio and video sequences. Consequently, our work stands as the inaugural and reference study directly tackling this distinctive and urgent issue.

In this context, we revisit the multimodal deep learning model presented in (Moufidi et al., 2023), originally created for biometrics task under short-utterances circumstances and that we adapt to Deepfake recognition. This innovative approach employs 2D decomposition to split a video into three distinct views, allowing an efficient spatiotemporal feature extraction at a minimal computational cost. To thoroughly validate our contributions, we conduct comprehensive evaluations across various benchmark datasets, thereby affirming the effectiveness of our model.

The remainder of this paper is structured as follows: Section 2 reviews existing literature in audio and visual deepfake detection, emphasizing the limitations and computational challenges of current approaches. Section 3 elaborates on our multimodal deep learning model for multimodal deepfake detection. Section 4 describes the experimental design, datasets used, and evaluation metrics. Section 5 presents a comprehensive analysis of our results, followed by Section 6 that concludes the paper and discusses future research directions.

2 RELATED WORKS

The realm of deepfake generation has witnessed the emergence of numerous techniques leveraging multiple modalities, including audio, visual, and their fusion (Masood et al., 2023). These networks for generating deepfakes have raised substantial concerns due to their potential use in spreading misinformation, manipulating public opinion, and fabricating identities. Consequently, researchers have responded by proposing a diverse array of countermeasures (Zi et al., 2020; Thing, 2023).

For instance, A. Pianese et al. (Cozzolino et al., 2023; Pianese et al., 2022) adopted a unique approach to address audio deepfake detection by leveraging a Person of Interest (POI) concept. This methodology essentially parallels the foundational principles of speaker verification systems. It seeks to measure the similarity between the voice being analyzed and a pre-existing reference set of the claimed identity. To achieve this, they incorporated two distinct non-supervised techniques: centroid-based testing and maximum-similarity testing (Cozzolino et al., 2023; Pianese et al., 2022). However, one inherent limitation of this approach is the necessity for a comprehensive reference set for each identity being analyzed.

In another strand of audio deepfake detection, researchers have experimented with handcrafted features like Mel Frequency Cepstral Coefficients (MFCC). These coefficients are processed through various 2D neural network architectures, such as VGG16 and EfficientNet, among others (Afchar et al., 2018; Chollet, 2017). A somewhat related approach has been to employ mel-spectrograms as features and feed them into transformer-based neural networks (Ilyas et al., 2023). While these methods have demonstrated promise, they are not without complexities, especially when compared to Time Delay Neural Networks (TDNN), which have shown to excel in tasks requiring temporal analysis (Desplanques et al., 2020).

Visual-based detection methods have seen a diverse range of strategies. Some leverage 3D networks for in-depth sequence analysis (Zhou and Lim, 2021; Zi et al., 2020), while others prioritize image-based methods with an emphasis on facial features (Ilyas et al., 2023; Zi et al., 2020). A noteworthy example is the late fusion architecture rooted in DST-Net by H. Ilyas et al. (Ilyas et al., 2023). Yet, even such advanced approaches frequently encounter difficulties in accurately analyzing ultra-short video sequences.

When it comes to visual modality, researchers have ventured into the use of 3D networks for sequence detection as well as image-based networks. These networks are designed to analyze the visual aspects of a deepfake, primarily focusing on facial features (Zi et al., 2020). Expanding the scope to multimodal systems, H. Ilyas et al. unveiled a late fusion architecture based on Dense Swin Transformer Net (DST-Net) (Ilyas et al., 2023). This intricate design extracts features separately from each modality—audio and visual—and then fuses them at the classification level to make a binary decision on the

realness of the media. The visual deepfake detection of the video was based on major voting, in other words, their 2D network analyses each frame from the sequence and based on the maximum classes gotten, the network decides if the video is fake or real. Unfortunately, this method does not take into account the temporal information and neither the possibility of the presence of deepfake short utterance in the video.

Adding another layer of complexity, Zhou et al. proposed a system that exploits the intrinsic synchronization between audio and visual elements, specifically focusing on the lips' movement and the corresponding audio. They employed a multimodal neural network and experimented with three types of fusion mechanisms based on attention mechanisms (Zhou and Lim, 2021). Despite its efficiency, this method presents a considerable computational burden due to the use of 3D networks and attention mechanisms, in addition, it is mainly destined for one language.

We extend the work (Moufidi et al., 2023), an architecture designed to mitigate the computational costs associated with the SOTA methods. This extension allows us to extract spatio-temporal features, thereby improving detection accuracy. This approach dissects videos into three distinct views: one spatial and two temporal.

3 METHODOLOGY

In scenarios with abundant data, deep learning, especially CNNs, proves invaluable due to its capability to autonomously discern and extract pivotal features, a feat often surpassing the performance of hand-crafted methods. Building on (Moufidi et al., 2023) developed deep learning model for biometric identification tasks, we have incorporated specific enhancements to address the unique challenges presented by deepfake detection in audio-visual data. These enhancements, detailed in the subsequent sections, refine the model's architecture and functionality, where we have changed the number of classes to 2 (Real Audio - Real Video and Fake Audio - Fake Video) or 4 (Real Audio - Real Video, Fake Audio - Fake Video, Real Audio - Fake Video and Fake Audio - Real Video), ensuring optimized performance for this application.

3.1 Audio-Visual Fusion

In our approach to deepfake detection, we opt for a late fusion technique inspired by the prior research in biometric tasks (Moufidi et al., 2023). This architecture is presented in Figure 1. The decision to employ late fusion is motivated by several factors. Firstly, it offers a more straightforward implementation compared to alternatives like early or hybrid fusion, effectively balancing information derived from both audio and visual modalities.

Secondly, the networks used for feature extraction in each modality are already pre-trained on expansive datasets. For audio, we use x-vectors trained on the VoxCeleb dataset (Chung et al., 2018), and for the visual aspect, we utilize ResNet18 trained on ImageNet (Deng et al., 2009). This allows us to concentrate solely on fine-tuning the fusion and classification layers, streamlining the overall training process.

Additionally, by dividing the video into multiple views, XY representing the spatial view, TY and XT incorporating the spatio-temporal respectively views across the y-axis and x- axis, we enhance our detection capabilities. This multi-view framework permits the identification of specific features such as jaw location thanks to the XY view, motion jitters TY view, and common artifacts that deepfake generators often struggle to simulate convincingly.

Lastly, the computational efficiency of a multi-view architecture makes it a more practical choice for feature extraction than using 3D CNNs, particularly in real-world applications.

4 EXPERIMENTS AND RESULTS

The primary focus of the current study is on the performance of the model under the constraints of short utterances. This evaluation is of particular importance given the increasing sophistication of deepfake generation techniques, which now have the capability to manipulate even brief segments of audio-visual data.

To evaluate (Moufidi et al., 2023) network, we exclusively consider videos from the classes R_vR_a (Real Video - Real Audio) and F_vF_a (Fake Video - Fake Audio) from FakeAvCeleb dataset (Khalid et al., 2021b), labeled respectively as real and fake audiovisual. These sequences are split to frame lengths ranging from $0.2ms$, $0.6ms$ and $1s$ with an overlap of 50%. We have ensured that there was an equal partition between fake and real labels. For all experiments, we only selected the lips part by using the Mediapipe tool offered by Google (Lugaresi et al., 2019).

The system's visual input consists of the lower facial region, selected for its computational efficiency. The 2D decomposition allows for an economical extraction of spatiotemporal information from the video footage. Despite this optimization, our approach yields superior detection accuracy relative to current state-of-the-art models, as indicated in Table 1.

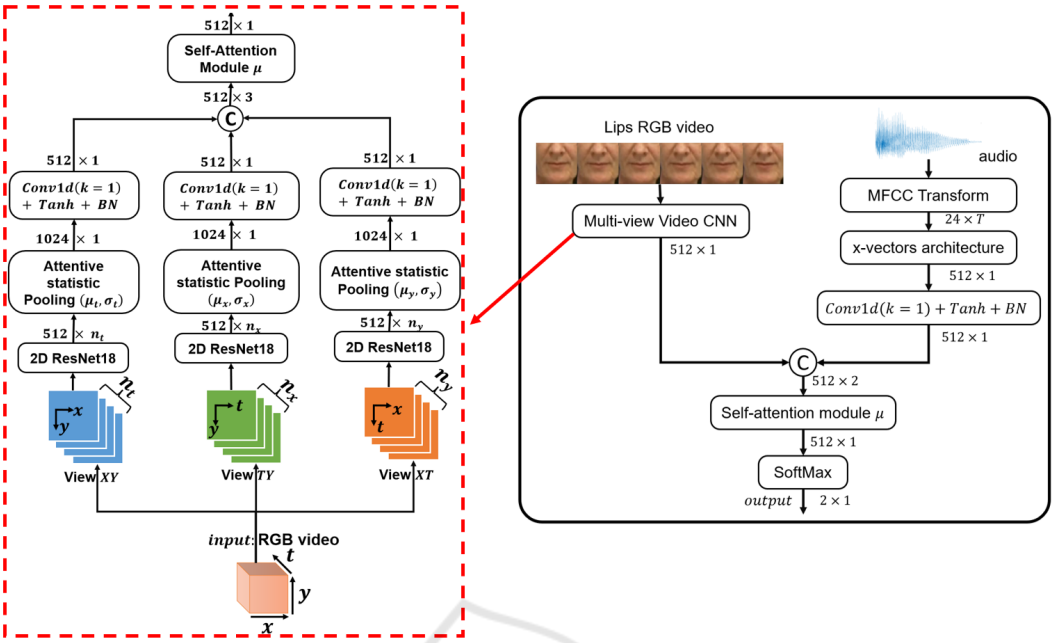

Figure 1: Multi-view CNN late fusion architecture for audio-lips correlation (Moufidi et al., 2023). For multimodal deepfake detection, the number of classes are set to 2 or 4, in addition, we only tune the fusion layers (all layers except 2D ResNet-18 and x-vectors).

Table 1: Comparison of SOTA performance with our proposed approach (Moufidi et al., 2023) on various time segments from FakeAVCeleb on balanced 1,000 samples (500 F_vF_a, 500 R_vR_a) subset, the train-test split was set to $80\% - 20\%$.

Model	**Accuracy (%)**		
	200 ms	**600 ms**	**1s**
XceptionNet (Soft-Voting) (Zi et al., 2020)	77.33	77.69	73.34
Multi-View CNN (Ours) (Moufidi et al., 2023)	93.19	97.68	98.55

4.1 Multi Class Detection

In this subsection, our emphasis is on evaluating our network's performance on short segments containing four distinct classes. The goal is to determine whether the network's performance remains consistent when at least one real modality is present, in other words, there are four main labels:

- Fake Video - Fake Audio (F_vF_a),
- Fake Video - Real Audio (F_vR_a),
- Real Video - Fake Audio (R_vF_a),
- Real Video - Real Audio (R_vR_a).

For this purpose, we selected a dataset comprising equally distribution among four subsets: F_vF_a, R_vF_a, F_vR_a and R_vR_a, token from FakeAVCeleb (Khalid et al., 2021b). We then segmented the lip portions of videos from these categories into frame lengths of $0.2s$, $0.6s$, and $1s$, with a 50% overlap between consecutive frames. The dataset was further divided into training and testing sets following an $80 - 20\%$ split, and the classification layer was set to 4 classes.

The results displayed in Table 3 confirm a huge decrease in detection accuracy compared to the scenario with only two classes, as outlined in Table 1. These observations indicate that the inclusion of a real modality tends to strongly influence our network's determination of a video's authenticity. To deeply understand the dominance of one modality over another, we carry out an experiment in the upcoming subsection to pinpoint the modality where our network demonstrates lower detection accuracy. Additionally, we explore the contribution of each view in improving the overall performance.

4.2 Investigation on the Role of Each View and Modality

The performance enhancements observed in our model can be attributed to a synergistic combination

Table 2: Detection accuracy of Multi-view CNN on short videos from FakeAVCeleb (train-test split $80\% - 20\%$ on $2,000$ videos of (500 F_vF_a, 500 F_vR_a, 500 R_vF_a and 500 R_vR_a), the classes number for classification is set to 4).

Window length	Accuracy (train/test)
200ms	85.95%/74.04%
600ms	82.89%/82.00%
1s	88.68%/85.96%

of view decomposition and the specific architecture employed for each modality. To delve deeper into the impact of each view—namely, XY, XT, TY —and their fusion, we conducted a dedicated experiment. For this evaluation, we considered a balanced dataset consisting of 500 real videos (R_vR_a) and 500 fake videos (F_vF_a), segmented into 1-second frame lengths with overlapping intervals. The data was partitioned into training and testing sets at an $80 - 20\%$ ratio, and the experiment was designed with two distinct classes.

The findings, presented in Table 3, illuminate the crucial roles played by individual views and modalities in the system's performance. Notably, the temporal view (TY), capable of accounting for issues like motion jitters, outshines the spatial (XY) and the spatiotemporal (XT) views. The relatively poorer performance of the spatial view equipped with a self-attention model can be explained by the focus on the lower part of the face—a known challenging aspect for many state-of-the-art deepfake detection methods, as cited in (Thing, 2023; Tolosana et al., 2022).

When fusing all visual views, we observe a marked increase in detection accuracy, thereby showcasing the model's prowess in resolving the ambiguity or confusion that could arise from individual views. Moreover, our results indicate that the audio modality holds a distinct edge in detection accuracy, contributing to an overall performance lift of 7.86% when integrated with the visual modality.

5 DISCUSSION

Navigating the intricate landscape of fake video detection necessitates a methodological framework that is both efficient and nuanced. In tuning the pre-existing model, cited as (Moufidi et al., 2023), we have not only successfully adapted it for fake video detection but also advanced our understanding of how different modalities contribute to the detection process. One of the primary strengths of the architecture lies in its utilization of pre-trained networks: ResNet-18 for ImageNet and x-vectors for VoxCeleb2. These well-established, data-rich training sources confer upon our model a robust initial feature set. Moreover, we adopt a computationally economical approach by using a 2D scheme for visual sequences and a 1D scheme for audio, thereby circumventing the need for more resource-intensive networks.

Our research goes beyond mere detection to dissect the relative contributions of each sequence view: XY, XT, and TY. The temporal motion jitters, belonging to TY view, emerges as the most accurate in detecting fake videos. This likely capitalizes on the inherent difficulties that deepfake algorithms have in accurately reproducing the temporal dynamics of human behavior. Conversely, the spatial view XY underperforms, which is consistent with existing literature (Thing, 2023; Tolosana et al., 2022) indicating that the lower facial region presents substantial challenges for deepfake detection systems.

The fusion of these three views adds an additional layer of complexity, further refining the model's detection capabilities. Such a fusion approach effectively exploits both spatial and temporal information, without the need for resource-intensive 3D models. Importantly, the incorporation of audio via x-vectors lends a significant boost to the model's performance. This may be attributed to the transfer learning advantages offered by VoxCeleb2, or it could point to a more fundamental characteristic of deepfake generation algorithms—that they are currently more proficient in visual manipulation than in audio.

Despite these promising outcomes, the architecture's performance is not without limitations. Most notably, its efficacy diminishes when applied to short utterances. This finding is significant and indicates a key area for future research: optimizing the model to maintain high detection rates irrespective of video length.

6 CONCLUSION

In this study, we have expanded upon the prior work in late fusion biometric recognition (Moufidi et al., 2023) to address the detection of deepfake videos using two distinct modalities. The model has demonstrated superior performance compared to the SOTA on the FakeAVCeleb dataset. Additionally, we have delved into the influence of three views in video decomposition and the role of modalities in augmenting detection accuracy. Notably, our findings highlight a substantial contribution from the audio modality in comparison to its visual counterpart.

Our future perspective aims to enhance this synergy by considering the joint detection framework

Table 3: Detection accuracy of Multi-view CNN on short videos from FakeAVCeleb (train and test on 1,000 videos of (500 F_vF_a, 500 R_vR_a) cutted into 1s frame length with an overlap 50%, the classes number for classification is set to 2).

Modality	View	Accuracy (train/test)	Precision (train/test)	Recall (train/test)
Visual	XY	85.76%/80.22%	88.12%/82.17%	89.64%/79.96%
Visual	XT	85.47%/83.04%	88.45%/87.68%	87.22%/84.93%
Visual	TY	88.15%/88.21%	91.31%/88.93%	88.72%/92.76%
Visual	XYT	94.88%/90.05%	96.21%/93.52%	95.23%/90.41%
Audio	–	100%/100%	100%/100%	100%/100%
Audio + Visual	XYT+ Audio	98.77%/97.91%	99.06%/97.96%	98.90%/98.72%

that could further integrate the audiovisual features at multiple levels of abstraction. Such a multimodal system could benefit from the inherent strengths of each modality, potentially leading to a more resilient detection mechanism against sophisticated deepfake manipulations. These efforts will contribute to the overarching goal of ensuring the authenticity and trustworthiness of digital media.

ACKNOWLEDGEMENTS

The authors thank Angers Loire Métropole (ALM) for the Ph.D grant of Abderrazzaq Moufidi.

REFERENCES

Afchar, D., Nozick, V., Yamagishi, J., and Echizen, I. (2018). Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258.

Chung, J. S., Nagrani, A., and Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*.

Cozzolino, D., Pianese, A., Nießner, M., and Verdoliva, L. (2023). Audio-visual person-of-interest deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 943–952.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Desplanques, B., Thienpondt, J., and Demuynck, K. (2020). Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*.

Huang, T.-h., Lin, J.-h., and Lee, H.-y. (2021). How far are we from robust voice conversion: A survey. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 514–521. IEEE.

Ilyas, H., Javed, A., and Malik, K. M. (2023). Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio-visual deepfakes detection. *Applied Soft Computing*, 136:110124.

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

Jiang, Z., Liu, J., Ren, Y., He, J., Zhang, C., Ye, Z., Wei, P., Wang, C., Yin, X., Ma, Z., et al. (2023). Mega-tts 2: Zero-shot text-to-speech with arbitrary length speech prompts. *arXiv preprint arXiv:2307.07218*.

Khalid, H., Kim, M., Tariq, S., and Woo, S. S. (2021a). 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.

Khalid, H., Tariq, S., Kim, M., and Woo, S. S. (2021b). Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*.

Korshunova, I., Shi, W., Dambre, J., and Theis, L. (2017). Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 3677–3685.

Ling, J., Tan, X., Chen, L., Li, R., Zhang, Y., Zhao, S., and Song, L. (2022). Stableface: Analyzing and improving motion stability for talking face generation.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M., Lee, J., Chang, W.-T., Hua, W., Georg, M., and Grundmann, M. (2019). Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR)*.

Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., and Malik, H. (2023). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence*, 53(4):3974–4026.

Moufidi, A., Rousseau, D., and Rasti, P. (2023). Attention-based fusion of ultrashort voice utterances and depth videos for multimodal person identification. *Sensors*, 23(13):5890.

Nirkin, Y., Keller, Y., and Hassner, T. (2019). Fsgan: Sub-

ject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7184–7193.

Pianese, A., Cozzolino, D., Poggi, G., and Verdoliva, L. (2022). Deepfake audio detection by speaker verification.

Prajwal, K., Mukhopadhyay, R., Namboodiri, V. P., and Jawahar, C. (2020). 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*, pages 484–492.

Sanderson, C. and Lovell, B. C. (2009). Multi-region probabilistic histograms for robust and scalable identity inference. In *Advances in Biometrics: Third International Conference, ICB 2009, Alghero, Italy, June 2-5, 2009. Proceedings 3*, pages 199–208. Springer.

Thing, V. L. (2023). Deepfake detection with deep learning: Convolutional neural networks versus transformers. *arXiv e-prints*, pages arXiv–2304.

Tolosana, R., Romero-Tapiador, S., Vera-Rodriguez, R., Gonzalez-Sosa, E., and Fierrez, J. (2022). Deepfakes detection across generations: Analysis of facial regions, fusion, and performance evaluation. *Engineering Applications of Artificial Intelligence*, 110:104673.

Zhang, T. (2022). Deepfake generation and detection, a survey. *Multimedia Tools and Applications*, 81(5):6259–6276.

Zhou, Y. and Lim, S.-N. (2021). Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14800–14809.

Zi, B., Chang, M., Chen, J., Ma, X., and Jiang, Y.-G. (2020). Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2382–2390.