Audio-Visual Deepfake Detection With Local Temporal Inconsistencies

Marcella Astrid⋆ *Enjie Ghorbel*⋆† *Djamila Aouada*⋆

⋆Interdisciplinary Centre for Security, Reliability and Trust (SnT),
University of Luxembourg, Luxembourg
†Cristal Laboratory, National School of Computer Sciences (ENSI),
Manouba University, Tunisia

ABSTRACT

This paper proposes an audio-visual deepfake detection approach that aims to capture fine-grained temporal inconsistencies between audio and visual modalities. To achieve this, both architectural and data synthesis strategies are introduced. From an architectural perspective, a temporal distance map, coupled with an attention mechanism, is designed to capture these inconsistencies while minimizing the impact of irrelevant temporal subsequences. Moreover, we explore novel pseudo-fake generation techniques to synthesize local inconsistencies. Our approach is evaluated against state-of-the-art methods using the DFDC and FakeAVCeleb datasets, demonstrating its effectiveness in detecting audio-visual deepfakes.

Index Terms— deepfake detection, audio-visual, fine-grained classification, augmentation

1. INTRODUCTION

Audio-visual deepfakes have beneficial applications but pose significant risks when misused [1, 2]. As their quality improves, distinguishing real from fake becomes increasingly difficult, highlighting the need for effective detection systems.

One approach for detecting audio-visual deepfakes is to identify inconsistencies between audio and visual data, as shown in numerous research works [3–6]. However, since deepfake artifacts are mostly subtle [7, 8], a fine-grained classification is needed. The use of a fine-grained approach in audio-visual deepfake detection has only recently been investigated in [9] by exploring both architecture design and data augmentation strategies. However, Astrid *et al.* [9] focus only on spatial inconsistencies, while exploring a very limited range of data augmentation techniques. To the best of our knowledge, no previous work has explored temporal fine-grained artifacts despite their informativeness in audio-visual deepfake detection. It can be noted that this line of work is distinct from fusion-based methods (e.g., [10–12]) that aim to fuse multimodal information, as well as from methods that extract the identity of individuals (e.g., [13, 14]).

This work was supported by the Luxembourg National Research Fund (FNR) under the project BRIDGES2021/IS/16353350/FaKeDeTeR and POST Luxembourg.

In this paper, we propose therefore to introduce mechanisms for modeling temporal fine-grained irregularities from both the architecture and the data synthesis perspectives. Specifically, in terms of architecture, we compute temporal distances for each time step between audio and visual data. These fine-grained temporal distances are fed into the classifier, as shown in Fig. 1(b). Our architecture also incorporates an attention mechanism to minimize the impact of irrelevant audio-visual distances, such as background sounds. Moreover, we propose extending the work of [9] where additional fake data are generated by replacing several frames from a given audio/video with data coming from a different audio/video, as shown in Fig. 1(a). In this work, to synthesise pseudo-fakes with subtle temporal artifacts, we suggest replacing sub-sequences from an original video with slightly manipulated versions of the same video using simple operations such as translation, flip, and frame repetition.

In summary, our contributions are as follows: 1) To the best of our knowledge, we are among the first to explore temporal local inconsistencies for detecting audio-visual deepfakes using both data augmentation and architecture strategies; 2) We propose augmenting the training data by generating pseudo-fake examples with subtle inconsistencies and exploring different methods for their generation; 3) We design a deepfake classifier that measures audio-visual feature distances across time to capture fine-grained temporal inconsistencies, enhanced with an attention mechanism; 4) We evaluate our method against state-of-the-art (SOTA) approaches using the DFDC dataset [15] for in-dataset and the FakeAVCeleb dataset [16] for cross-dataset settings.

Paper organization. Our methodology is detailed in Section 2, while the experiments and the results are detailed in Section 3. Finally, the conclusion is given in Section 4.

2. METHODOLOGY

As shown in Fig. 1, we address the fine-grained deepfake detection problem at the temporal level from two perspectives, namely: (1) data augmentation with pseudo-fake generation (Section 2.1); and (2) architectural design for capturing fine-grained audio-visual distances in the temporal dimension (Section 2.2).

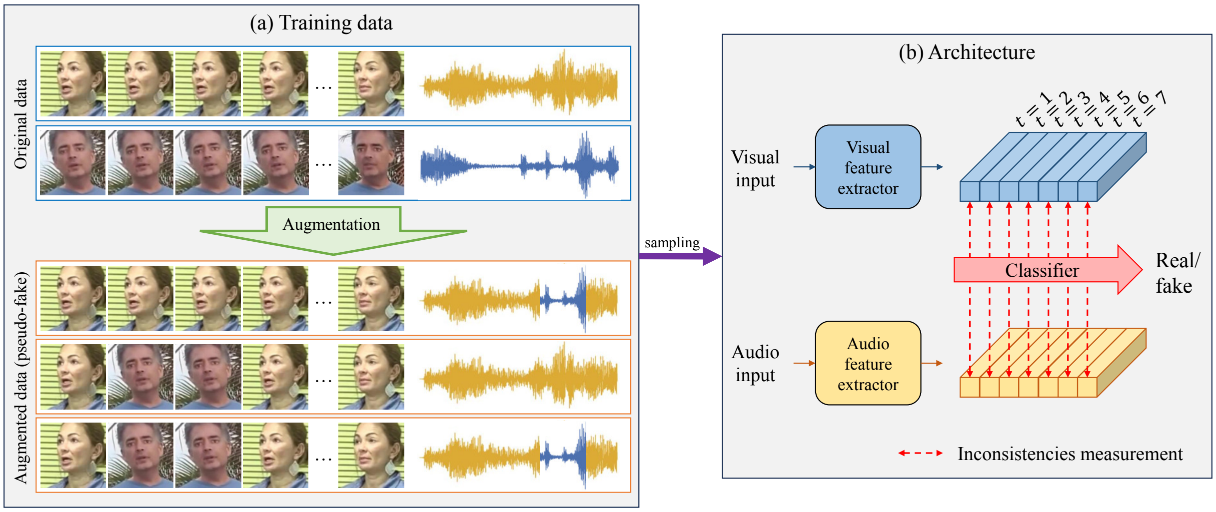

Fig. 1. We address temporal fine-grained inconsistencies from (a) the data synthesis perspective and (b) the architectural perspective. (a) We augment the original data with pseudo-fake examples by locally modifying the data in the temporal domain. (b) The classifier evaluates audio-visual inconsistencies at each temporal step.

2.1. Data augmentation

The original training set consists of pairs of audio-visual data, $\mathbf{X}^v$ and $\mathbf{X}^a$, with sizes $T^v \times C^v \times H^v \times W^v$ and $T^a \times C^a$, respectively. Here, the superscripts v and a denote visual and audio data, respectively. T, C, H, and W represent time, channel, height, and width dimensions, respectively. Since we use a waveform input for the audio, C^a equals 1. The audio and video sequences are extracted from the same part of the overall video.

We generate pseudo-fakes from visual $\mathbf{P}^v$ and audio $\mathbf{P}^a$ data to augment the original training set. Each pseudo-fake data point is randomly selected as either $\{\mathbf{X}^v, \mathbf{P}^a\}$, $\{\mathbf{P}^v, \mathbf{X}^a\}$, or $\{\mathbf{P}^v, \mathbf{P}^a\}$, and is labeled as *fake*. We use pseudo-fake data as input instead of the original data with a probability of 0.5.

We generate $\mathbf{P}^v$ and $\mathbf{P}^a$ in a similar way. For the sake of simplicity, we use a shared variable $\mathbf{A} = \{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_T\}$ to represent both $\mathbf{X}^v$ and $\mathbf{X}^a$, with $T = T^v$ for visual data and $T = T^a$ for audio data. Each $\mathbf{A}_i$ represents a frame of size $C^v \times H^v \times W^v$ for visual data or a waveform magnitude value for audio data. The input $\mathbf{A}$ is then manipulated to create the pseudo-fake.

2.1.1. Temporally-local manipulation

Following [9], to create a pseudo-fake $\mathbf{P}$ with subtle artifacts from $\mathbf{A}$, we locally modify the sequence from the time step i to $i + l - 1$. We denote this modified chunk as $\mathbf{C} = \mathbf{A}_{i:i+l-1} = \{\mathbf{A}_i, \mathbf{A}_{i+1}, \ldots, \mathbf{A}_{i+l-1}\}$. For example, in Fig. 2(a), i and l are 3 and 4, respectively, so $\mathbf{C} = \{\mathbf{A}_3, \mathbf{A}_4, \mathbf{A}_5, \mathbf{A}_6\}$. The modification length l is randomly selected within $[l_{min}, l_{max}]$, where

$$l_{min} = r_{min} \times T, \qquad l_{max} = r_{max} \times T, \tag{1}$$

with r_{min} and r_{max} being hyperparameters in the range $]0, 1]$. We set the minimum l_{min} to 2.

2.1.2. Manipulation types

There can be several ways to modify $\mathbf{C}$ so that the resulting sequence differs from the original $\mathbf{A}$:

1. Replacing with another clip

$$\mathbf{C} = \mathbf{B}_{i:i+l-1}, \tag{2}$$

where $\mathbf{B}_{i:i+l-1}$ is a subsequence from a randomly selected data sample. This method, depicted in Fig. 2(b) is used in [9]. In this work, we also consider other methods.

2. Repeating

$$\mathbf{C} = \mathbf{A}_{i+\lfloor \frac{a}{p} \rfloor * p} \quad \text{where } a = \{0, 1, ..., l-1\}, \tag{3}$$

where p is the number of repetitions, randomly chosen from 2 to l for each generated pseudo-fake. In the example of Fig. 2(c), we use $p = 2$.

3. Flipping

$$\mathbf{C} = \mathbf{A}_{i+2f\lfloor \frac{a}{f} \rfloor + f - 1 - a} \quad \text{where } a = \{0, 1, ..., l-1\}, \tag{4}$$

where f is the flipping frequency, randomly chosen from 2 to l for each pseudo-fake. In the example of Fig. 2(d), we uses $f = 2$.

4. Translating left or right

$$\mathbf{C} = \mathbf{A}_{i+\min(l-1, a+v)} \text{ where } a = \{0, 1, ..., l-1\}, \text{ or} \tag{5}$$

$$\mathbf{C} = \mathbf{A}_{i+\max(0, a-v)} \quad \text{where } a = \{0, 1, ..., l-1\}, \tag{6}$$

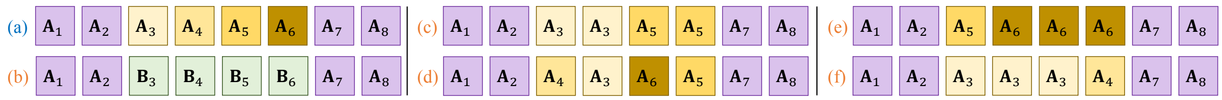

Fig. 2. Given (a) an original data sequence, we generate pseudo-fake data by locally modifying the sequence. In this example, we modify $\mathbf{A}_3$ to $\mathbf{A}_6$. We explore various modifications: (b) replacing with sub-sequences from another data sample, (c) repeating, (d) flipping, and (e-f) translating from the left or right.

where v is the translation step, randomly selected from 2 to l. The direction, left or right, is also chosen randomly. Examples in Fig. 2(e) and (f) use $v = 2$.

2.2. Architectural design

To capture subtle temporal inconsistencies, we design a classifier that processes the distance map at each time step. The architecture, shown in Fig. 3, includes feature extractors, distance calculation, an attention mechanism, and a classifier.

2.2.1. Feature extractors

Given a pair of audio-visual inputs $\mathbf{X}^v$ (or $\mathbf{P}^v$) and $\mathbf{X}^a$ (or $\mathbf{P}^a$), we extract features as follows:

$$\mathbf{F}^v = \mathcal{F}^v(\mathbf{X}^v), \qquad \mathbf{F}^a = \mathcal{F}^a(\mathbf{X}^a), \tag{7}$$

where $\mathbf{F}^v$ and $\mathbf{F}^a$ are features of size $T' \times C'$, with T' and C' representing the time and channel dimensions. The visual feature extractor $\mathcal{F}^v(\cdot)$ is a ResNet-based 3D Convolution (Conv3D) model, designed to be shallow to preserve a larger temporal dimension. We then use a 3D adaptive average pooling block to output features with a spatial size of 1 and a temporal dimension equal to T'. The audio feature extractor $\mathcal{F}^a(\cdot)$ is based on 1D Convolution (Conv1D) layers to produce features of size $T' \times C'$.

2.2.2. Temporal local distance

To identify fine-grained temporal inconsistencies, we compute a distance map $\mathbf{m}$ of size T', where each element represents the L2 distance between the features of the two modalities:

$$m_t = ||\mathbf{f}_t^v - \mathbf{f}_t^a||, \tag{8}$$

where $\mathbf{f}_t^v$ and $\mathbf{f}_t^a$ are the features $\mathbf{F}^v$ and $\mathbf{F}^a$ at an instant t, respectively.

2.2.3. Attention mechanism

To filter out irrelevant audio, such as background noise, we add an attention mechanism to focus on important pairs. We compute an attention map $\mathbf{a}$ of the same size as $\mathbf{m}$ using a cross-attention mechanism. First, we calculate an intermediate vector $\mathbf{a}'$ such that,

$$a'_t = \left(\frac{\mathcal{E}^a(\mathbf{F}^a)_t \cdot \mathcal{E}^v(\mathbf{F}^v)_t}{C'} \right), \tag{9}$$

where $\mathcal{E}^v(\cdot)$ and $\mathcal{E}^a(\cdot)$ are trainable 1D convolutions with $C'/4$ filters to reduce computation, while maintaining the dimension T'. We then use softmax to compute the final attention map:

$$a_t = \frac{\exp(a'_t)}{\sum_{j=1}^{T'} \exp(a'_t)}. \tag{10}$$

This attention map shows the degree of correlation between the audio and visual features, hence reducing the impact of unrelated noise.

Finally, the attended distance map $\hat{\mathbf{m}}$ is calculated as,

$$\hat{\mathbf{m}} = \mathbf{m} \odot \mathbf{a}, \tag{11}$$

where $\odot$ represents the element-wise product. It can be noted that a similar attention mechanism is used in [9], but applied to the temporal rather than the spatial distance.

2.2.4. Classifier

The classifier $\mathcal{C}$ processes $\hat{\mathbf{m}}$ to predict the probability that the input is fake, y:

$$y = \mathcal{C}(\hat{\mathbf{m}}). \tag{12}$$

The model is trained using a binary cross-entropy loss.

3. EXPERIMENTS

3.1. Experimental setup

Dataset. Following prior works [3, 9, 17], we train on a subset of the DFDC dataset [15] with 15,300 training and 2,700 test videos, balancing real and fake videos in the training set while maintaining the original test set distribution (*in-dataset* protocol). To evaluate the generalization capability of the model, we test on the FakeAVCeleb dataset [16] as *cross-dataset* evaluation, using 70 real and 70 fake videos, as in [18]. We follow the data processing of [3], except that we use the waveform format for audio, normalized between -1 to 1. The final dimensions are $T^v = 30$, $C^v = 3$, $H^v = W^v = 224$, $T^a = 48000$.
Metric. We evaluate the performance of our method using the video-level Area Under the ROC Curve (AUC). The video-level prediction is calculated by averaging the prediction of the subsequences.
Implementation details. We train the model for 50 epochs using the Adam optimizer [19] with a learning rate of 10^{-3}, weight decay of 10^{-5}, and a batch size of 8, evaluating based on the lowest training loss. Unless specified otherwise, we set $T' = 15$, $C' = 128$, $r_{min} \approx 0$ (i.e., $l_{min} = 2$), and $r_{max} = 1$.

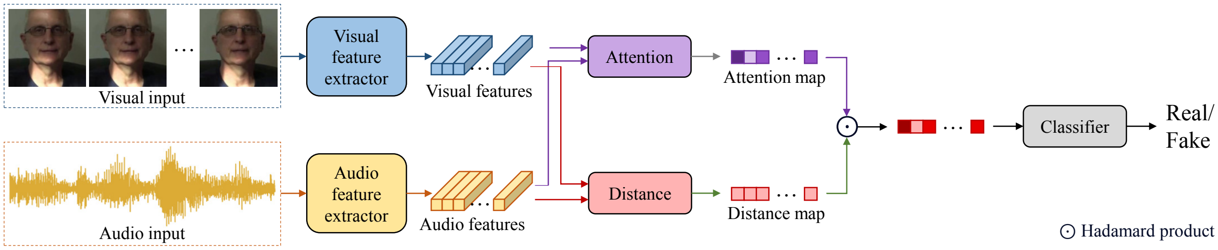

Fig. 3. Our architecture uses a fine-grained distance map for each time step of the extracted features, combined with an attention mechanism, to classify whether the input pair is fake.

	Type	AUC (In-dataset)	AUC (Cross-dataset)
(a)	No augment	96.42%	60.98%
(b)	Replacing	96.76%	**80.34%**
(c)	Repeating	97.01%	67.44%
(d)	Flipping	**97.24%**	70.16%
(e)	Translating	96.65%	62.04%

Table 1. Comparison in terms of AUC of different manipulation techniques used to create pseudo-fakes, as discussed in Section 2.1.2.

	Temporal size	AUC (In-dataset)	AUC (Cross-dataset)
(a)	1	87.20%	67.61%
(b)	7	**98.03%**	**87.02%**
(c)	15	96.76%	80.34%

Table 2. Model comparison in terms of AUC with varying T' values.

3.2. Ablation study

In this subsection, we evaluate the importance of each component and compare different component configurations.

3.2.1. Pseudo-fake manipulation

We compare manipulation techniques from Section 2.1.2. As shown in Table 1(b)-(e), all outperform the no-augmentation baseline (Table 1(a)) in both in-dataset and cross-dataset settings. Replacing with another clip performs best, likely because it mirrors content replacement in deepfake generation. Hence, we adopt this technique for the subsequent experiments.

3.2.2. Attention

The importance of the attention mechanism in our method is demonstrated by comparing the model performance with and without it. The model incorporating attention achieves an AUC of 96.76% in the in-dataset setting and 80.34% in the cross-dataset setting, outperforming the model without attention, with an AUC of 96.63% and 72.14%, respectively. This improvement highlights the effectiveness of the attention mechanism in enhancing the model's performance.

3.2.3. Distance map size

Table 2 compares models with different T' values. The $T' = 1$ setup (Table 2(a)) performs worse than the ones set with higher T' values, indicating that fine-grained distances are more effective than global distances (Table 2(b)-(c)). Interestingly, a very high T', such as $T' = 15$, can slightly reduce the performance compared to $T' = 7$. This may be due to the model becoming too sensitive and detecting inconsistencies in low-quality real videos.

Method	AUC	Method	AUC
MDS [3]	90.7%	VFD [13]	85.1%
Emotion [17]	84.4%	AvoiD-DF [10]	94.8%
BA-TFD [14]	84.6%	SADD [6]	96.7%
AVT^2-DWF [11]	89.2%	FGI [9]	97.7%
AVFakeNet [12]	86.2%	Ours	**98.0%**

Table 3. Comparison with SOTA in terms of AUC on DFDC (in-dataset).

Method	AUC	Method	AUC
MDS [3]	72.9%	SADD [6]	61.4%
AVoiD-DF [10]	82.8%	FGI [9]	84.5%
AVT^2-DWF [11]	77.2%	Ours	**87.0%**

Table 4. Comparison with SOTA in terms of AUC on FakeAVCeleb (cross-dataset).

3.3. Comparison to the SOTA

We compare our method with state-of-the-art (SOTA) audio-visual deepfake detection approaches, using the optimal setup ($T' = 7$ with attention and clip replacement as an augmentation method). Our method outperforms SOTA in both in-dataset (Table 3) and cross-dataset settings (Table 4). The improved performance over FGI [9] can suggest that the temporal local distance is more suitable than spatial local distance for detecting audio-visual deepfakes.

4. CONCLUSION

We propose detecting audio-visual deepfakes by identifying temporal inconsistencies. Our approach includes both data augmentation and architectural design strategies. For the augmentation, we experiment various manipulation techniques to create pseudo-fakes. As for the architecture, we assess the impact of local versus global distances and the role of attention mechanisms. Our method surpasses state-of-the-art approaches under both in-dataset (DFDC) and cross-dataset (FakeAVCeleb) settings.

5. REFERENCES

[1] Heather Chen and Kathleen Magramo, "Finance worker pays out $25 million after video call with deepfake 'chief financial officer'," *CNN*.

[2] Gianluca Mezzofiore, "Deepfake video targeting zelensky's wife linked to russian disinformation campaign, cnn analysis shows," *CNN*.

[3] Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian, "Not made for each other-audio-visual dissonance-based deepfake detection and localization," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 439–447.

[4] Yewei Gu, Xianfeng Zhao, Chen Gong, and Xiaowei Yi, "Deepfake video detection using audio-visual consistency," in *Digital Forensics and Watermarking: 19th International Workshop, IWDW 2020, Melbourne, VIC, Australia, November 25–27, 2020, Revised Selected Papers 19*. Springer, 2021, pp. 168–180.

[5] 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*, 2023, pp. 10491–10503.

[6] Marcella Astrid, Enjie Ghorbel, and Djamila Aouada, "Statistics-aware audio-visual deepfake detector," *arXiv preprint arXiv:2407.11650*, 2024.

[7] Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada, "Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17395–17405.

[8] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu, "Multi-attentional deepfake detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2185–2194.

[9] Marcella Astrid, Enjie Ghorbel, and Djamila Aouada, "Detecting audio-visual deepfakes with fine-grained inconsistencies," *arXiv preprint arXiv:2408.06753*, 2024.

[10] 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*, vol. 18, pp. 2015–2029, 2023.

[11] Rui Wang, Dengpan Ye, Long Tang, Yunming Zhang, and Jiacheng Deng, "Avt2-dwf: Improving deepfake detection with audio-visual fusion and dynamic weighting strategies," *arXiv preprint arXiv:2403.14974*, 2024.

[12] Hafsa Ilyas, Ali Javed, and Khalid Mahmood Malik, "Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio–visual deepfakes detection," *Applied Soft Computing*, vol. 136, pp. 110124, 2023.

[13] Harry Cheng, Yangyang Guo, Tianyi Wang, Qi Li, Xiaojun Chang, and Liqiang Nie, "Voice-face homogeneity tells deepfake," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023.

[14] 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)*. IEEE, 2022, pp. 1–10.

[15] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer, "The deepfake detection challenge (dfdc) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[16] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo, "Fakeavceleb: A novel audio-video multimodal deepfake dataset," in *Proc. Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track*, 2021.

[17] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha, "Emotions don't lie: An audio-visual deepfake detection method using affective cues," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 2823–2832.

[18] 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*, 2021.

[19] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.