RawBMamba: End-to-End Bidirectional State Space Model for Audio Deepfake Detection

Yujie Chen[1], Jiangyan Yi[2,], Jun Xue[1], Chenglong Wang[2], Xiaohui Zhang[2], Shunbo Dong[1], Siding Zeng[2], Jianhua Tao[3], Lv Zhao[1], Cunhang Fan[1]*

[1] School of Computer Science and Technology, Anhui University, China [2] Institute of Automation, Chinese Academy of Sciences, China [3] Department of Automation, Tsinghua University, China

e22201148@stu.ahu.edu.cn, jiangyan.yi@nlpr.ia.ac.cn

Abstract

Fake artefacts for discriminating between bonafide and fake audio can exist in both short- and long-range segments. Therefore, combining local and global feature information can effectively discriminate between bonafide and fake audio. This paper proposes an end-to-end bidirectional state space model, named RawBMamba, to capture both short- and long-range discriminative information for audio deepfake detection. Specifically, we use sinc Layer and multiple convolutional layers to capture short-range features, and then design a bidirectional Mamba to address Mamba's unidirectional modelling problem and further capture long-range feature information. Moreover, we develop a bidirectional fusion module to integrate embeddings, enhancing audio context representation and combining short- and long-range information. The results show that our proposed RawBMamba achieves a 34.1% improvement over Rawformer on ASVspoof2021 LA dataset, and demonstrates competitive performance on other datasets. Codes will be released on https://github.com/cyjie429/RawBMamba.

Index Terms: state space model, bidirectional mamba, audio deepfake detection

1. Introduction

Recently, advances in deep learning have significantly accelerated the development of text-to-speech (TTS) and voice conversion (VC) technologies, making synthesised speech increasingly indistinguishable from real human voices. Nevertheless, the potential misuse of such technologies by malicious actors inevitably poses a significant threat to society. Hence, in recent years, a series of fake audio detection challenges [1, 2, 3, 4] emerge with the aim of countering and bolstering the ability to detect fake audio attacks.

Extensive research demonstrates that discriminative information (i.e., fake artefacts) of synthesized and converted speech can be analyzed from both short-range and long-range segments of speech signals [5, 6]. Abnormal intonation changes and unnatural stresses may exist in short-range segments, while unnatural rhythmic patterns and monotonous emotional expressions as deceptive cues may be present in long-range segments. Therefore, integrating both short-range and long-range information effectively enhances the system's detection capability. In recent years, mainstream audio deepfake detection approaches have been divided into pipeline [7, 8, 9] and end-to-end methods. The end-to-end model [10, 11, 12, 13], by not relying on manually extracted features and directly optimizing on the raw audio waveform within a single model, can effectively enhance the model's generalizability. Inspired by this, we believe that extracting both short- and long-range feature information from audio signals through an end-to-end approach can ensure the model's effectiveness and generalizability. Currently, extensive research focuses on end-to-end models, RawNet2 [14, 15] by performing time-domain convolution operations directly on raw audio, possesses the potential to learn clues undetectable by knowledge-based methods. Additionally, researchers incorporate attention mechanisms into graph neural networks to capture key information across time and frequency [16, 17, 18]. AASIST [5] captures complex short-range and long-range feature information in graph networks by modeling the non-Euclidean relationships of graph nodes. Rawformer [6] integrates convolutional layers with Transformer [19] structures, utilizing self-attention to capture short-range and long-range feature information in speech signals. Although the aforementioned end-to-end models are already capable of handling audio deepfake detection, there still exists room for improvement.

Recently, inspired by the success of the state space model with efficient hardware-aware designs (i.e., Mamba [20]) in the field of sequence modeling, which selects relevant information to model long-range relationships, empirical evidence has demonstrated its enhanced capability in capturing long-range relations more effectively. Therefore, we believe it can also more effectively capture long-range feature information in speech signals. Based on this, we aim to design an efficient end-to-end model that integrates both short-range and long-range feature information. However, Mamba, due to its Markovian nature, suffers from the drawback of unidirectional modeling, which prevents it from fully capturing contextual information, thereby affecting the acquisition of long-range information.

To address these limitations, we propose a End-to-End Bidirectional Mamba model, named RawBMamba, to more effectively capture short- to long-range feature information. Specifically, we utilize a series of parametrizable sinc functions to obtain low-level feature mappings, followed by acquiring high-level feature mappings through multiple convolutional blocks, followed by mapping reconstruction to generate a two-dimensional sequence suitable for bidirectional Mamba. The Mamba models are then used to capture forward and backward long-range features respectively. Finally, a bidirectional feature fusion module merges the two sets of embeddings to obtain the comprehensive short-range to long-range feature representation. In addition, we perform a preliminary theoretical analysis of the effectiveness of Mamba and Transformers in the domain of audio deepfake detection to demonstrate the viability of Mamba. We achieve competitive experimental results on multiple datasets, proving the effectiveness and generalization ability of RawBMamba.
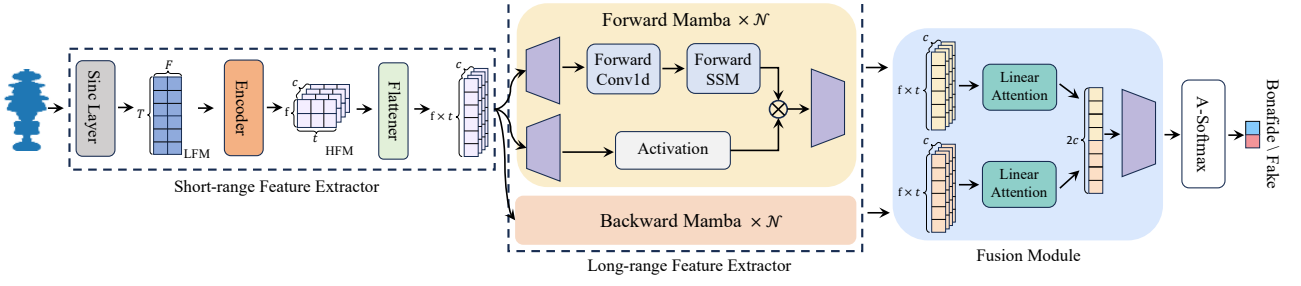
---

* Corresponding author.

Figure 1: *The overall structure diagram of our proposed RawBMamba.*

2. Preliminaries

Mamba is inspired by continuous state space models (SSM) [21, 22, 23, 24] in control systems. It maps a 1-D function or sequence $x(t) \in \mathbb{R} \rightarrow y(t) \in \mathbb{R}$ through a hidden state $h(t) \in \mathbb{R}^N$. This model uses $\mathbf{A} \in \mathbb{R}^{N \times N}$ as the State Transition parameters, and $\mathbf{B} \in \mathbb{R}^{N \times 1}$, $\mathbf{C} \in \mathbb{R}^{1 \times N}$ as the projection parameters.

$$
\begin{aligned}
h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\
y(t) &= \mathbf{C}h(t).
\end{aligned} \tag{1}
$$

Mamba is the classic discrete versions of the continuous system. it include a timescale parameter $\mathbf{\Delta}$, and through discretization rules, transform the continuous parameters $\mathbf{A}$, $\mathbf{B}$ to discrete parameters $\overline{\mathbf{A}}$, $\overline{\mathbf{B}}$. A commonly used discretization rule is the zero-order hold (ZOH), which is defined as follows:

$$
\begin{aligned}
\overline{\mathbf{A}} &= \exp(\mathbf{\Delta A}), \\
\overline{\mathbf{B}} &= (\mathbf{\Delta A})^{-1}(\exp(\mathbf{\Delta A}) - \mathbf{I}) \cdot \Delta\mathbf{B}.
\end{aligned} \tag{2}
$$

After the discretization of $\overline{\mathbf{A}}$, $\overline{\mathbf{B}}$, the discretized version of Eq. (1) using a step size $\mathbf{\Delta}$ can be rewritten as:

$$
\begin{aligned}
h_t &= \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t, \\
y_t &= \mathbf{C}h_t.
\end{aligned} \tag{3}
$$

At last, the models compute output through a global convolution.

$$
\begin{aligned}
\overline{\mathbf{K}} &= (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}}\overline{\mathbf{B}}, \ldots, \mathbf{C}\overline{\mathbf{A}}^{M-1}\overline{\mathbf{B}}), \\
y &= x * \overline{\mathbf{K}}.
\end{aligned} \tag{4}
$$

where M is the length of the input sequence $\mathbf{x}$, and $\overline{\mathbf{K}} \in \mathbb{R}^M$ is a structured convolutional kernel.

3. Proposed Method

In this section, we provide a detailed introduction to RawB-Mamba, as shown in Figure 1 . First, we describe how RawB-Mamba captures short-range feature representation. Then, we present the implementation details of the Bidirectional Mamba and the Bidirectional Feature Fusion Module.

3.1. Short-range feature representation

An increasing number of researchers are using trainable neural layers to learn approximate standard filter processes from raw waveforms. RawBMamba employs a variant of RawNet2's frontend to capture high-level short-range feature mappings

---

Algorithm 1 Bidirectional Mamba

Input: token sequence $F_S : (B, L, C)$
Output: token sequence $F_{forward}, F_{backward} : (B, L, C)$
  /* sequence reverse operation */
  $F_{forward} : (B, L, C) \leftarrow F_{short}$
  $F_{backward} : (B, L, C) \leftarrow \text{Reverse}(F_{short})$
  /* process with different direction */
  for d in {forward, backward} **do**
    $x : (B, L, C) \leftarrow F_d$
    for i in $MambaLayer_d$ **do**
      $x_i, z : (B, L, E) \leftarrow \text{Linear}^{xz}(x))$
      $x_i' : (B, L, E) \leftarrow \text{SiLU}(\text{Conv1d}(x_i))$
      /* Selection Mechanism */
      $\mathbf{B}_i : (B, L, N) \leftarrow \text{Linear}_i^B(x_i')$
      $\mathbf{C}_i : (B, L, N) \leftarrow \text{Linear}_i^C(x_i')$
      /* softplus ensures positive Δ_i */
      $\mathbf{\Delta}_i : (B, L, E) \leftarrow \log(1 + \exp(\text{Linear}_i^\Delta(x_i') + \text{Parameter}_i^\Delta))$
      /* shape of Parameter$_i^\mathbf{A}$ is (E, N) */
      $\overline{\mathbf{A}_i} : (B, L, E, N) \leftarrow \mathbf{\Delta}_i \otimes \text{Parameter}_i^\mathbf{A}$
      $\overline{\mathbf{B}_i} : (B, L, E, N) \leftarrow \mathbf{\Delta}_i \otimes \mathbf{B}_i$
      $y_i : (B, L, C) \leftarrow \text{Linear}(\text{SSM}(\overline{\mathbf{A}_i}) \otimes \text{SiLU}(z))$
      $x : (B, L, C) \leftarrow y_i$
    end for
    $F_d : (B, L, C) \leftarrow y_i$
  end for
  return $F_{forward}, F_{backward}$

(HFM). Specifically, first, a series of parametric sinc functions are utilized to implement band-pass filters [25], extracting spectral-temporal features from the raw waveform, forming Low-level Feature Maps (LFM) $F_{LFM} \in \mathbb{R}^{F \times T}$, where where F and T are the numbers of frequency and the temporal bins, respectively. Then, High-level Feature Maps (HFM) $F_{HFM} \in \mathbb{R}^{C \times f \times t}$ containing short-range feature information are extracted through a series of three ResNet blocks with squeeze-and-excitation operations, here, C, f and t denote the number of channels, frequency bins, and temporal locations after dimensionality reduction. Finally, the high-level feature maps are flattened along the time and frequency axes to obtain a two-dimensional short-range feature sequence $F_s \in \mathbb{R}^{C \times ft}$, suitable as input vectors for the bidirectional Mamba.

3.2. Bidirectional state space model

Mamba designs a simple selection mechanism that, by parameterizing the sequence dimension, enables the model to effectively select the most relevant information and filter out irrele-

vant information, thereby enhancing the model's ability to capture long-range features. However, due to its Markovian nature, it can only model long-range features in a unidirectional manner, which leads to ineffective capture of contextual information. Therefore, we design the bidirectional Mamba, as illustrated in Figure 1, which specifically consists of the following two parts:

Bidirectional Mamba. We present Algorithm 1 for the bidirectional Mamba. we first take the short-range features F_s obtained from the frontend and generate a backward short-range feature through a reverse sequence operation. Then, we deploy two structurally identical Mamba networks to independently capture the long-range feature information of the forward and backward short-range feature. Specifically, we project the forward and backward features onto x_i and z with dimensions of E, respectively. Then a 1-D convolution is applied to x to obtain x_i'. Following that, x_i' is projected to obtain $\mathbf{B}_i$, $\mathbf{C}_i$ and $\mathbf{\Delta}_i$. Discretisation rules are used to generate $\overline{\mathbf{A}}_i$, $\overline{\mathbf{B}}_i$. Finally, the forward long-range features $F_{forward}$ and the backward long-range features $F_{backward}$ are obtained.

Bidirectional feature fusion block. After acquiring the forward and backward long-range features, we initially apply a linear self-attention operation on each of the unidirectional long-range features to extract key information. Subsequently, we employ concatenation for bidirectional feature fusion. Finally, we obtain short- and long-range speech features F_l enriched with contextual information, which are used for authenticity discrimination. The specific formula is illustrated as follows:

$$F_1, F_2 = Attention(F_{forward}), Attention(F_{backward}) \tag{5}$$

$$F_l = MLP(Concat(F_1, F_2)) \tag{6}$$

4. Experiments and Analysis

4.1. Datasets and metrics

We conduct extensive tests on the ASVspoof2019 LA (19LA), ASVspoof2021 LA (21LA) and ASVspoof2021 DF (21DF) to evaluate our model's effectiveness and generalizability. 19LA features three types of fake attacks (TTS, VC) across 19 algorithms (A01-A19). 21LA contains real and fake speeches via telephony systems like Voice over IP and Public Switched Telephone Network. 21DF comprises authentic and spoofed voices altered by various media codecs, which introduces distortion during the encoding, compressing, and subsequent decoding of audio data. The performance assessment uses the equal error rate (EER) and the minimum tandem detection cost function (min t-DCF).

4.2. Implementation details

In the training phase, we use input data with 64,000 sampling points ($\approx$ 4 seconds) for RawBMamba. The SincNet layer is configured with 70 filters, and the variant based on RawNet2 consists of four sub-blocks, with the first two having 32 filters and the latter two having 64 filters each. We utilize an Adam optimizer with a learning rate of 1×10^{-5}, and training batch size is set to 32. The model is trained on the ASVspoof 2019 LA training and development sets for 32 epochs using the A-Softmax[26] loss function on a single RTX 3090 GPU.

Table 1: *The experimental results of RawBMamba with different configurations on the 19LA, 21LA, and 21DF datasets. Here, "uni" is the unidirectional Mamba and "bi" is the bidirectional Mamba. "L" is model's total layers.*

Dir.	L	19LA		21LA		21DF
		EER(%)	t-DCF	EER(%)	t-DCF	EER(%)
uni	4	2.51	0.0707	3.74	0.2707	**20.94**
	8	2.01	0.0623	3.44	0.2652	21.22
	12	**1.47**	**0.0467**	**2.84**	**0.2517**	22.48
bi	4	**1.07**	**0.0315**	3.38	0.2687	20.83
	8	1.67	0.0484	3.45	**0.2631**	18.80
	12	1.19	0.0360	**3.28**	0.2709	**15.85**

Table 2: *The experimental results of RawBMamba with different fusion methods on the 19LA, 21LA, and 21DF datasets.*

Methods	19LA		21LA		21DF
	EER(%)	t-DCF	EER(%)	t-DCF	EER(%)
Sum	1.27	0.0400	4.13	0.2924	16.58
Concat	**1.19**	**0.0360**	3.28	0.2709	**15.85**
Attention	1.19	0.0369	**3.19**	**0.2620**	18.42

4.3. Results of RawBMamba with different configurations

This section discusses the unidirectional and bidirectional Mamba models' capability to capture short- and long-range features with N layers, as detailed in Table 1. We observe that the unidirectional Mamba model struggles with the 21DF dataset and only shows effectiveness on 21LA, indicating limitations in capturing generalized long-range features. However, when analysing comprehensive results from all datasets, the bidirectional Mamba model consistently outperforms the unidirectional Mamba model. This shows that bidirectional Mamba can overcome the problem of insufficient contextual information capture in unidirectional Mamba, thereby improving the generalizability of long-range features. In addition, we observe that the bidirectional Mamba model maintains good performance even with a lower number of layers (i.e., $N = 4$) compared to unidirectional Mamba models with fewer layers, which we attribute to the effectiveness of our bidirectional feature fusion module in fully exploiting information from both forward and backward long-range features. However, the reduction in the number of layers inevitably leads to a reduction in the ability to handle difficult datasets such as 21DF.

4.4. Results of RawBMamba with different fusion methods

In this section, we will discuss in detail the impact of different fusion methods in the bidirectional feature fusion module of RawBMamba, as shown in Table 2. Specifically, we design three common fusion methods: summation, concatenation, and the cross-attention mechanism. Both summation and cross-attention mechanism act directly on the two unidirectional long-range features. The results show that concatenation achieves the best overall effect compared to the other two fusion methods, maintaining good performance across all datasets. This suggests that while preserving the integrity of the time-frequency information as much as possible, it is able to capture the key information of the two unidirectional long-range features in the

Table 3: *The comparison results of RawBMamba, Rawformer, and AASIST on the 19LA, 21LA, and 21DF datasets. Here, "†" is reproduced results.*

Models	19LA		21LA		21DF
	EER(%)	t-DCF	EER(%)	t-DCF	EER(%)
T23 [2]	-	-	1.32	0.2177	15.64
RawNet2 [14]	-	-	9.50	-	22.38
TO-RawBet [27]	1.58	-	3.70	-	-
AASIST [5]	**0.93**	**0.0285**	10.51	0.4884	-
AASIST-4Block [5]	1.20	0.0341	9.15	0.4370	-
ARawNet2 [28]	4.61	-	8.36	-	19.03
SE-Rawformer [6]	1.05	0.0344	4.98	0.3186	-
SE-Rawformer†	1.15	0.0314	4.31	0.2851	20.26
RawMamba(ours)	1.47	0.0467	**2.84**	**0.2517**	22.48
RawBMamba(ours)	1.19	0.0360	3.28	0.2709	**15.85**

time-frequency domain. The cross-attention mechanism performs poorly on the 21DF dataset, suggesting that it may over-focus on key information and lose time-frequency information, reducing generalisation to out-of-domain data. In contrast, the summation fusion method shows promising results on the 21DF dataset, as it involves a simple summation operation on all time-frequency domain information of the two unidirectional long-range features, acting as a potential regularization effect that balances the information across all time-frequency domains.

4.5. Comparison with the other end-to-end models

In Table 3, we compare the performance of RawBMamba on 19LA, 21LA, and 21DF with the other end-to-end models. RawMamba refers to a 12-layer unidirectional Mamba model, whereas RawBMamba is a 12-layer bidirectional Mamba model, with 6 layers per direction.

To verify RawBMamba's generalization capability, We conduct extensive experiments on three datasets, with RawB-Mamba consistently performing well across all. Specifically, on the 21LA dataset, our proposed RawBMamba significantly outperforms the other two baselines. RawBMamba shows a 34.1% performance improvement over SE-Rawformer, and RawMamba shows a 43.0% improvement over SE-Rawformer, with comparable results on the 19LA dataset. This indicates that Mamba effectively captures long-range feature information more efficiently than Transformer. Furthermore, experimental results on the 21DF dataset show that RawBMamba, as a single system, achieves performance extremely close to that of the multi-system score fusion T23. This indicates that RawB-Mamba maintains commendable performance even when faced with complex, diverse, and unknown speech forgery methods. It demonstrates RawBMamba's robustness to out-of-domain data, further proving its generalization capability.

4.6. Effectiveness analysis of RawBMamba and Rawformer

We discuss whether the Mamba architecture has more potential than the Transformer architecture for handling the audio fake detection task. Specifically, we directly extract the final features on the 19LA test dataset using the Rawformer model and the RawBMamba model. Then, we use t-SNE [29], a non-linear dimensionality reduction algorithm for visualizing high-dimensional data, to visualize the features. The results are shown in Figure 2. From (a) and (b), it is not difficult to see
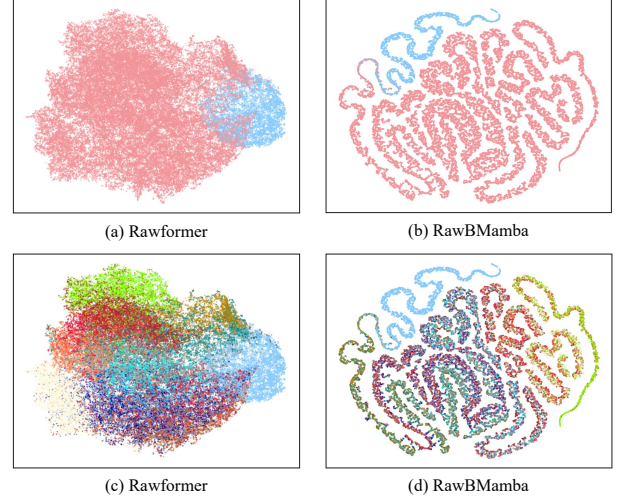

Figure 2: *The 19LA test set samples' clustering is shown in 2D t-SNE plots from the model's higher layers. The visualization displays the clustering of bonafide vs. fake audio in (a) and (b) (blue for bonafide, red for fake) and different attack types in (c) and (d), with each color indicating a specific attack.*

that when dealing with the binary task of bonafide and fake audio detection, there is partial block overlap between bonafide and fake audio for Transformer, while Mamba only has thread-like overlap at the end of the curve-shaped figures, indicating that the features obtained by Mamba are more discriminable than those extracted by Transformer. Furthermore, we also discover that Mamba's feature visualization presents multiple non-overlapping curvilinear patterns, embodying richer feature information capable of distinguishing different attack methods. This indicates Mamba architecture's potential in handling fine-grained classification tasks, as shown in (c) and (d). When mapping the sample points of different attack methods, the distribution of Mamba is clearer, which We believe is due to the key role of Mamba's sequence selection mechanism in obtaining discriminative information between different attack methods.

Therefore, we believe the Mamba architecture is capable of handling audio deepfake detection and has significant potential to become a backbone in audio deepfake detection models.

5. Conclusion

In this paper, we propose an end-to-end state space model, named RawBMamba, to capture both short- and long-range discriminative information in audio signals. Specifically, We use parametrizable sincLayers and multiple convolutional layers to capture short-range features, then design a bidirectional Mamba to address the limitations of Mamba's unidirectional modeling. We use a bidirectional feature fusion module to merge forward and backward long-range features, enhancing audio context representation. Experiments indicate that RawBMamba achieves a 34.1% improvement over the Rawformer on ASVspoof 2021 LA and outperforms the state-of-the-art end-to-end models on ASVspoof 2021 LA, validating RawBMamba's generalizability and capability to handle out-of-domain data. Visualization analysis on t-SNE plots further validates the effectiveness of RawB-Mamba and its potential in fine-grained classification tasks. In future work, we plan to explore the application of RawBMamba to provenance tasks.

6. Acknowledgements

7. References

[1] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "Asvspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.

[2] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans *et al.*, "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge*, 2021.

[3] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan *et al.*, "Add 2022: the first audio deep synthesis detection challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9216–9220.

[4] J. Yi, J. Tao, R. Fu, X. Yan, C. Wang, T. Wang, C. Y. Zhang, X. Zhang, Y. Zhao, Y. Ren *et al.*, "Add 2023: the second audio deepfake detection challenge," *arXiv preprint arXiv:2305.13774*, 2023.

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

[6] X. Liu, M. Liu, L. Wang, K. A. Lee, H. Zhang, and J. Dang, "Leveraging positional-related local-global dependency for synthetic speech detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

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

[8] C. Wang, J. Yi, J. Tao, C. Y. Zhang, S. Zhang, and X. Chen, "Detection of Cross-Dataset Fake Audio Based on Prosodic and Pronunciation Features," in *Proc. INTERSPEECH 2023*, 2023, pp. 3844–3848.

[9] X. Zhang, J. Yi, J. Tao, C. Wang, and C. Y. Zhang, "Do you remember? overcoming catastrophic forgetting for fake audio detection," in *International Conference on Machine Learning, ICML 2023*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 41 819–41 831.

[10] X. Li, X. Wu, H. Lu, X. Liu, and H. Meng, "Channel-wise gated res2net: Towards robust detection of synthetic speech attacks," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association 2021*. ISCA, 2021, pp. 4314–4318.

[11] A. Chintha, B. Thai, S. J. Sohrawardi, K. Bhatt, A. Hickerson, M. Wright, and R. Ptucha, "Recurrent convolutional structures for audio spoof and video deepfake detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 1024–1037, 2020.

[12] G. Hua, A. B. J. Teoh, and H. Zhang, "Towards end-to-end synthetic speech detection," *IEEE Signal Processing Letters*, vol. 28, pp. 1265–1269, 2021.

[13] Y. Ma, Z. Ren, and S. Xu, "RW-Resnet: A Novel Speech Anti-Spoofing Model Using Raw Waveform," in *Proc. Interspeech 2021*, 2021, pp. 4144–4148.

[14] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6369–6373.

[15] J. H. Hansen and Z. WANG, "Audio anti-spoofing using simple attention module and joint optimization based on additive angular margin loss and meta-learning," in *Proc. Interspeech 2022*, 2022, pp. 376–380.

[16] H. Tak, J. Jung, J. Patino, M. Todisco, and N. W. D. Evans, "Graph attention networks for anti-spoofing," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*. ISCA, 2021, pp. 2356–2360.

[17] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[18] H. Tak, J.-W. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," in *ASVSPOOF 2021, Automatic Speaker Verification and Spoofing Countermeasures Challenge*. ISCA, 2021, pp. 1–8.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[21] A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, and C. Ré, "Combining recurrent, convolutional, and continuous-time models with linear state space layers," *Advances in neural information processing systems*, vol. 34, pp. 572–585, 2021.

[22] A. Gu, K. Goel, A. Gupta, and C. Ré, "On the parameterization and initialization of diagonal state space models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 35 971–35 983, 2022.

[23] A. Gupta, A. Gu, and J. Berant, "Diagonal state spaces are as effective as structured state spaces," *Advances in Neural Information Processing Systems*, vol. 35, pp. 22 982–22 994, 2022.

[24] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. [Online]. Available: https://openreview.net/forum?id=uYLFoz1vlAC

[25] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE spoken language technology workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[26] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[27] C. Wang, J. Yi, J. Tao, C. Y. Zhang, S. Zhang, R. Fu, and X. Chen, "TO-Rawnet: Improving RawNet with TCN and Orthogonal Regularization for Fake Audio Detection," in *Proc. INTERSPEECH 2023*, 2023, pp. 3137–3141.

[28] J. Li, Y. Long, Y. Li, and D. Xu, "Advanced rawnet2 with attention-based channel masking for synthetic speech detection," in *Proc. INTERSPEECH 2023*, 2023, pp. 2788–2792.

[29] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.