The Role of Long-Term Dependency in Synthetic Speech Detection

Changtao Li, Feiran Yang, *Member, IEEE*, and Jun Yang, *Senior Member, IEEE*

***Abstract*—Although much progress has been made in synthetic speech detection, there lacks comprehensive analysis of the essential differences between spoofed and genuine speech. We here utilize supervised contrastive loss originated from contrastive learning as an analytical tool to characterize the class similarity structure of ASVspoof 2019 logical access (LA) dataset, which shows that an ideal back-end classifier for synthetic speech detection should have the ability to capture long-term dependencies. Recently, Transformer has been found to have an excellent ability in learning long-term dependencies of input data. We hence propose a back-end classifier based on Transformer Encoder for synthetic speech detection. Convolution blocks are added before the Transformer Encoder, which leverages inductive biases to improve the generalization ability. Compared to two-dimensional convolution, one-dimensional convolution makes better architectural assumptions about the input speech features, which helps with modeling long-term dependencies and decreases the risk of overfitting. The proposed Transformer combined with one-dimensional convolution has fewer parameters than most existing back-end classifiers, and achieves an equal error rate of 1.06% and a minimum tandem detection cost function metric of 0.0345 when evaluated on ASVspoof 2019 LA dataset, which is one of the best models reported in the literature.**

***Index Terms*—ASVspoof 2019 LA, generalization ability, speaker verification, transformer, voice anti-spoofing.**

I. Introduction

THE recent development of voice conversion (VC) and text-to-speech (TTS) brings challenges to automatic speaker verification (ASV) systems [1], [2]. The spoofing detection built before ASV systems is an effective countermeasure. To foster progress in the development of spoofing detection systems, the ASVspoof organization has been providing series of datasets since 2015 [3]–[5]. Among those datasets, the logical access (LA) subset of ASVspoof 2019 dataset is the most used dataset for developing and benchmarking countermeasures against VC and TTS attacks [6].

While several end-to-end spoofing detection systems have been proposed [7]–[9], most spoofing detection systems consist of a front-end feature extractor and a back-end binary classifier. Back-end classifiers based on advanced architectures such as residual-connected convolution neural network (CNN) and light CNN have been proven successful when combined with handcrafted front-end features, e.g., linear frequency cepstral coefficient (LFCC) and constant Q cepstral coefficients (CQCC) [10]–[24]. Nevertheless, back-end classifiers based on deep learning models continue to be treated as black-box function approximators. As pointed out by Muller *et al.* [25], the back-end may pick up on data artifacts that are not related to spoofing attacks, which is counterproductive for research and development purpose. Therefore, it is necessary to do some analysis about spoofed speech so that a proper back-end can be chosen.

In this paper, we utilize supervised contrastive loss (SCL) originated from contrastive learning as an analytical tool to investigate the class similarity structure of the LFCC features of speech utterances in ASVspoof 2019 LA dataset [26]. We show that the class similarity decreases as the length of LFCC increases, and hence deep learning architectures that can capture long-term dependencies are highly preferred for synthetic speech detection.

Transformers are found to outperform CNNs and recurrent neural networks (RNNs) in learning long-term dependencies [27], [28]. Therefore, we propose to employ Transformer Encoder as the backbone of the back-end classifier in synthetic speech detection. However, due to the lack of inductive biases such as spatial locality and translation equivariance [29], Transformer Encoder usually requires much more training data than CNNs. A Transformer trained from scratch on ASVspoof 2019 LA training set is not guaranteed to generalize well. To remedy this limitation, we add several convolution blocks before Transformer Encoder, which leverages the inductive biases inherent to CNNs and helps to improve the generalization ability. We finally evaluate the performance of the proposed approach on ASVspoof 2019 LA dataset. The contributions of this paper are twofold. First, we propose one-dimensional convolutional Transformer (OCT) which performs better than the original Transformer in synthetic speech detection. Second, we show that experimental methods like sequence pooling (SeqPool) can further improve the generalization ability of our OCT model.

Manuscript received March 11, 2022; accepted April 7, 2022. Date of publication April 25, 2022; date of current version May 11, 2022. This work was supported in part by the National Natural Science Foundation of China under Grants 62171438 and 11974376, in part by the Youth Innovation Promotion Association of Chinese Academy of Sciences under Grant 2018027, and in part by IACAS Frontier Exploration Project under Grant QYTS202111. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiaodong Cui. *(Corresponding authors: Feiran Yang; Jun Yang.)*

Changtao Li and Jun Yang are with the Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: lichangtao@mail.ioa.ac.cn; jyang@mail.ioa.ac.cn).

Feiran Yang is with the State Key Laboratory of Acoustics, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China (e-mail: feirany.ioa@gmail.com).

Digital Object Identifier 10.1109/LSP.2022.3169954

TABLE I
SUMMARY OF ASVSPOOF 2019 LA DATASET

	#Genuine utterances	#Spoofed utterances	Attacks types
Training	2580	22800	A01-A06
Developing	2548	22296	A01-A06
Evaluating	7355	63882	A07-A19

II. ANALYSIS OF SPOOFED SPEECH

Though remarkable progress has been made [30]–[32], data generation techniques have limitations. For instance, the state-of-the-art image generation models excel at generating small images with few structural constraints, but they fail to capture the geometric or structural patterns that occur consistently in images [33]. In light of the close relationship between audio generation and image generation models [34], [35], the similar problems are likely to appear in TTS and VC.

It is possible that inconsistent global structural patterns are inherent to spoofed speech and its frame-level features. In other words, while local pieces of spoofed speech can be similar to these of genuine speech, the overall spoofed speech can exhibit obvious defects. To illustrate this concern, the class entanglement between utterances needs to be calculated.

A. Class Similarity and SCL

Objective functions for contrastive learning can measure the entanglement of class manifolds in representation space. The entanglement of different classes can reflect how close pairs of data points from the same class are relative to pairs of data points from different classes [36]. We here adopt the SCL presented by Prannay *et al.* [26] as the objective function. Given a dataset $\{\mathbf{z}_s\}, s \in I \equiv \{1, \ldots, N\}$, each element in this dataset has a corresponding label $y_s \in \{1, \ldots, L\}$ among L classes. The SCL is defined as

$$L_{out}^{sup} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)} \quad (1)$$

where $\cdot$ denotes the inner dot product, $A(i) \equiv I \backslash \{i\}$, $P(i) \equiv \{p \in A(i) : y_p = y_i\}$, $|P(i)|$ is the cardinality of $P(i)$, and τ is the temperature hyperparameter.

A large SCL means a strong entanglement of class manifolds. Specifically, data points from different classes are similar to each other when a dataset has a large SCL [36].

B. The Relationship Between SCL and Speech Length

ASVspoof 2019 LA dataset is composed of genuine speech and different kinds of spoofing attacks (A01–A19) [37]. Table I presents a summary of ASVspoof 2019 LA dataset. We investigate utterances in this dataset to reveal some common features of spoofed speech. To verify our aforementioned argument, the class entanglement between utterances as a function of utterance length needs to be exploited. The entanglement between front-end features would provide more information for designing back-end since raw audio waveform is not directly fed into the classifier. In what follows, we will show how to obtain the class entanglement between LFCCs as a function of the utterance length.

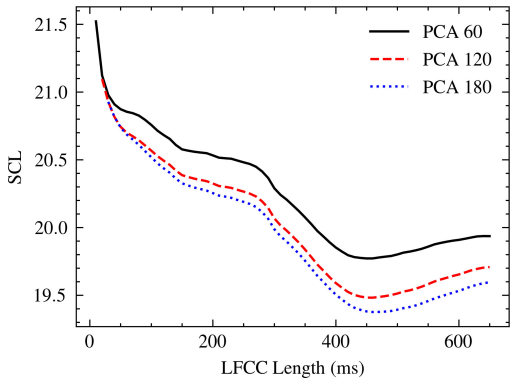

Fig. 1. SCL as a function of the LFCC length under different PCA dimensions.

To proceed, we produce a family of datasets α_s, $s \in 1, \ldots, N_{max}$ from ASVspoof 2019 LA training set. Each dataset α_s consists of LFCC features with the same frame number F_s, and F_s is monotonically increasing as s. Then, we can get the relationship between class entanglement and the length of LFCC by computing SCL for each α_s.

We show how to obtain α_s and compute the SCL for α_s.

1) We select 2580 spoofed utterances from the LA training set randomly. Those spoofed utterances, along with 2580 genuine utterances in LA training set, constitute a base dataset.
2) We compute the 60-dimensional LFCC feature of each speech clip in base dataset $\mathcal{L}_i \in \mathbb{R}^{T_i \times 60}$, $i \in 1, \ldots, 5160$, where T_i is the total frame number of $\mathcal{L}_i$. The frame length is 20 ms, and the hop length is 10 ms. Because the length of each speech clip varies, T_i is different for each i.
3) For each $\mathcal{L}_i$, $i \in 1, \ldots, 5160$, F_s contiguous frames are selected randomly and are then flattened into a one-dimensional tensor $P_i \in \mathbb{R}^{60F_s}$, $i \in 1, \ldots, 5160$.
4) We project P_i into $\widetilde{P}_i \in \mathbb{R}^E$ with principal component analysis (PCA), where E is the preset projection dimension.
5) We compute the SCL of $\{\widetilde{P}_i, i \in 1, \ldots, 5160\}$ using (1).

Fig. 1 shows the SCL values[1] as a function of the length of LFCC, where we set the projection dimension $E = 60, 120, 180$, respectively. Observed from Fig. 1 that the class entanglement decreases as F_s increases, which implies that a short piece of spoofed speech can be very similar with genuine speech. This also complies with the intuition that it is easier to synthesize shorter utterances. The SCL curve reaches its minimum at 450 ms, i.e., 45 frames. This length is much bigger than the kernel size of typical CNNs and may cause issues when updating the parameters of RNNs using gradient descent [29]. Therefore, CNNs and RNNs may not be the best backbone for the back-end, and classifiers that can better model long-term dependencies should be adopted for synthetic speech detection. The high entanglement between one-frame LFCCs also indicates that information along frequency-axis of LFCC may be redundant for detecting well-developed attacks considered in ASVspoof 2019 LA dataset. However, the frequency information may be useful in other cases.

[1] τ is kept the same for all three PCA dimensions. P_i is not normalized before PCA.

III. One Dimension Convolutional Transformer

Self-attention mechanism has been shown to outperform CNNs and RNNs on capturing the long-term dependencies of input data [27], [28]. Transformer Encoder based purely on self-attention may be a good candidate for the back-end classifier. However, we have already found that Transformer models trained without sufficient data generally lack good generalization capability. To apply Transformer in spoofing detection, we propose OCT that combines Transformer Encoder with one-dimensional convolution blocks.

A. Transformer Encoder

We use conventional Transformer Encoder, which follows the same structure as that in original Transformer [27] and vision Transformer (ViT) [28]. The Encoder is composed of multiple Transformer blocks, each including an attention and a position-wise feed-forward sub-block. Both attention and feed-forward sub-block comprise layer norm and residual connection. The output $\mathbf{k}_o$ of a sub-block (attention or feed-forward) with input sequence $\mathbf{k}$ is

$$\mathbf{k}_o = LayerNorm(\mathbf{k} + F(\mathbf{k})) \tag{2}$$

where F could be either a multi-head self-attention (MSA) layer or a position-wise feed-forward network.

B. Convolutional Blocks and Positional Embedding

The efficiency of convolution operation is mainly attributed to three important features, i.e., sparse interaction, weight sharing and equivariant representations [29]. However, these properties are missing in Transformers. As pointed out by Dosovitskiy *et al.* [28], Transformers do not generalize well when trained on insufficient amounts of data due to the lack of some inductive bias inherent to CNNs. Hence, the generalization ability can be enhanced by integrating Transformer Encoder with additional convolutional blocks. As validated by Ali *et al.*, the introduced two-dimensional convolutional blocks enable the Transformer variant (CCT) to generalize well on small datasets including CIFAR-10 and CIFAR-100 [38]. A similar strategy could be used to improve the generalization ability of Transformer in our system. Here, we use one-dimensional convolution instead of the two-dimensional one. As analyzed before, information along the frequency-axis of LFCC feature may not contribute too much in spoofing detection. Therefore, one-dimensional convolution is utilized to ignore redundant patterns along frequency-axis, which will decrease the risk of overfitting.

The architecture of convolutional block consists of a single one-dimensional convolution layer, ReLU activation, and a max pool operation. The input of the block is a tensor $\mathbf{x}$ with size of $C \times T$, where C and T represent the number of time steps and channels respectively. The output of this block is

$$\mathbf{x}_o = MaxPool(ReLU(Conv1d(\mathbf{x}))) \tag{3}$$

Learnable positional embedding is used to embed the position information of the output.

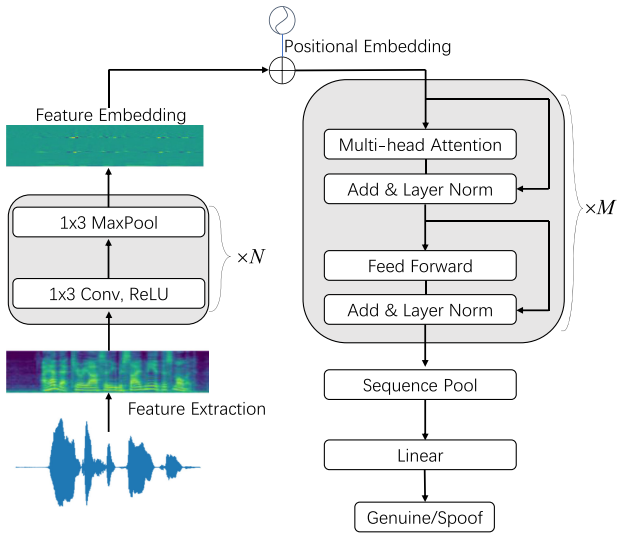

Fig. 2. Illustration of the overall structure of our proposed OCT.

C. SeqPool and Classification Head

We use SeqPool instead of the special $[CLS]$ token [28], [39] to pool the sequential output of Transformer Encoder $\mathbf{x}_L$. SeqPool is a mapping transformation [38]:

$$T : \mathbf{x}_L \in \mathbb{R}^{L \times D} \rightarrow \mathbf{z} \in \mathbb{R}^{D} \tag{4}$$

where $\mathbf{z}$ is the feature for the final binary classification, L is the sequence length, and D is the embedding dimension of the Transformer Encoder. SeqPool comprises three operations. First, $\mathbf{x}_L$ is fed into a linear layer g with one unit, and softmax activation is applied to get the weight for each element in this sequence:

$$\mathbf{w} = softmax\left(g(\mathbf{x}_L)^T\right) \in \mathbb{R}^{1 \times L} \tag{5}$$

Second, the sequential output $\mathbf{x}_L$ is averaged with these weights $\mathbf{w}$:

$$\mathbf{z} = \mathbf{w}\mathbf{x}_L = softmax\left(g(\mathbf{x}_L)^T\right) \times \mathbf{x}_L \in \mathbb{R}^{1 \times D} \tag{6}$$

Finally, the output of SeqPool $\mathbf{z} \in \mathbb{R}^D$ is obtained by squeezing the original tensor $\mathbf{z}$.

SeqPool yields the utterance representation from the sequential output of Transformer Encoder. For spoofing detection, the final output layer is a linear layer with two units, which represent the confidence score for each class.

D. Small Model Size of OCT

Fig. 2 shows the overall structure of our OCT model, where all the Conv layers apply "SAME" padding. The one-dimensional CNN block and the Transformer Encoder layer are repeatedly used for N and M times. The choice of $N \leq 3, M \leq 3$ is preferred because abstract semantic information is not the main focus of spoofing detection [7]. Additionlly, the number of heads of MSA layers should also be small to avoid memorizing irrelevant information along frequency-axis.

TABLE II
HYPERPARAMETERS OF USED OCT

	Specific layers	Input-Output shape
Tokenizer	Conv1d(64, 3, 1), ReLU(), MaxPool(3, 2)	(60, 512)→(64, 256)
	Conv1d(64, 3, 1), ReLU(), MaxPool(3, 2)	(64, 256)→(64, 128)
	Conv1d(128, 3, 1), ReLU(), MaxPool(3, 2)	(64, 128)→(128, 64)
	Transpose	(128, 64)→(64, 128)
Transformer Classifier	Transformer Encoder	(64, 128)→(64, 128)
	Transformer Encoder	(64, 128)→(64, 128)
	SeqPool	(64, 128)→(128)
	Linear	(128)→(2)

TABLE III
COMPARISION BETWEEN OCT, CCT AND ORIGINAL TRANSFORMER

Systems	t-DCF	EER
OCT	0.0345	1.06
CCT	0.0617	2.31
Original Transformer	0.0660	2.62

TABLE IV
OCT WITH DIFFERENT POOLING LAYERS

Systems	t-DCF	EER
SeqPool	0.0345	1.06
$[CLS]$ Token	0.0516	2.11
Average Pooling	0.0655	2.43

TABLE V
PERFORMACE COMPARISION

Systems	Front-end	t-DCF	EER	#Params
RawGAT-ST(mul) [8]	waveform	0.0335	1.06	0.44M
OCT	LFCC	0.0345	1.06	0.25M
Res-TSSDNet [7]	waveform	0.0481	1.64	0.35M
Raw PCDARTS Mel_f [9]	waveform	0.0517	1.77	24.48M
MCG-Res2Net50+CE [22]	CQT	0.0520	1.78	-
ResNet18-LCML-FM [14]	LFB	0.0520	1.81	-
LCNN-LSTM-sum [20]	LFCC	0.0524	1.92	0.28M

IV. EXPERIMENTS

We now evaluate the performance of the proposed model on ASVspoof 2019 LA dataset. Equal error rate (EER) and tandem detection cost function (t-DCF) are utilized as the performance indicators. A smaller EER means better performance of spoofing detection. While EER only focuses on the performance of spoofing detection, t-DCF considers the influence of spoofing detection on the reliability of an ASV system. A lower value of t-DCF implies a better reliability of ASV [40].

A. Training Details

We extract 60-dimensional LFCCs from the utterances. The frame size is 20 ms, and the hop length is 10 ms. Each LFCC feature is cropped or padded to 512 frames so that mini-batch can be formed. We use focal loss [41] with $\alpha = 0.75$ and $\gamma = 2.0$ as the objective function. The proposed OCT is used as back-end classifier, which takes the LFCC features as input and outputs the class probability. The hyperparameters of OCT are presented in Table II. Parameters in Conv1d layer and Maxpool layer refer to (number of filters, filter size, stride) and (filter size, stride), respectively. The output sizes before and after Transpose operation refer to (CNN channels, time) and (time, embedding dimension).

The proposed model is implemented by PyTorch. We use AdamW optimizer with learning rate 8e-4 and learning rate decay 1e-4. The mini-batch size is 64. All above hyperparameters are optimized based on model's performance on validation subset. After hyperparameter fine-tuning, we merge the training and validation subset and then train our model for 300 epochs on this merged set using one Nvidia RTX 3090 GPU. The converged model on this merged dataset is adopted for the evaluation. All other models mentioned in this paper are also fine-tuned and evaluated with this strategy.

B. Results

To show the importance of the existence of CNNs, we implement a classifier based on the original Transformer which has no convolution. Table III compares the performance of OCT, CCT and original Transformer. SeqPool is used in all three models. It is clear from Table III that all models achieve promising detection performance in this task. However, OCT significantly outperforms CCT and original Transformer in terms of both EER and t-DCF, which indicates that patterns along frequency-axis learned by two-dimensional convolution may not contribute to spoofing detection but increases the risk of overfitting. This validates our analysis about the LFCC feature of spoofed speech and demonstrates the merits of one-dimensional convolution.

We then study the effect of SeqPool on the detection performance. We replace the SeqPool operation in our OCT model with the special $[CLS]$ token pooling and a simple average pooling, and present the result in Table IV. It is clear that SeqPool is useful for further improving the generalization.

We finally compare our model with competing models reported in the literature in Table V. The proposed OCT significantly outperforms LCNN-LSTM-sum model [20] that also uses the LFCC feature. Therefore, it is apparent that Transformer Encoder is superior for spoofing detection compared to LSTM. The proposed model also achieves a comparable performance with the best end-to-end model [8]. Additionally, our system only requires 0.25 M parameters and is the least complex model so far.

V. CONCLUSION

In this paper, we analyzed the LFCC features of spoofed utterances in ASVspoof 2019 LA training set, which shows that long-term dependencies play an important role for spoofing detection. We hence proposed OCT that can learn long-term dependencies of spoofed speech and generalize well on ASVspoof 2019 LA dataset. With only 0.25 M parameters, the proposed model outperforms all back-end classifiers reported in the literature to date and its performance is comparable to the best end-to-end models on ASVspoof 2019 LA dataset.

REFERENCES

[1] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, "Advances in anti-spoofing: From the perspective of ASVspoof challenges," *APSIPA Trans. Signal Inf. Process.*, vol. 9, 2020, Art. no. e2.

[2] R. K. Das *et al.*, "Predictions of subjective ratings and spoofing assessments of voice conversion challenge 2020 submissions," in *Proc. Joint Workshop Blizzard Challenge Voice Convers Challenge*, 2020, pp. 99–120.

[3] Z. Wu *et al.*, "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 2037–2041.

[4] T. Kinnunen *et al.*, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 2–6.

[5] M. Todisco *et al.*, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 1008–1012.

[6] A. Nautsch *et al.*, "ASVspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Trans. Biometrics Behav. Identity Sci.*, vol. 3, no. 2, pp. 252–265, Apr. 2021.

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

[8] H. Tak, J. 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 *Proc. ASVspoof Workshop*, 2021, pp. 1–8.

[9] W. Ge, J. Patino, M. Todisco, and N. Evans, "Raw differentiable architecture search for speech deepfake and spoofing detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 4319–4323.

[10] J. Sanchez, I. Saratxaga, I. Hernáez, E. Navas, D. Erro, and T. Raitio, "Toward a universal synthetic speech spoofing detection using phase information," *IEEE Trans. Inf. Forensics Secur.*, vol. 10, no. 4, pp. 810–820, Apr. 2015.

[11] T. B. Patel and H. A. Patil, "Combining evidences from MEL cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 2062–2066.

[12] C. Zhang, C. Yu, and J. H. L. Hansen, "An investigation of deep-learning frameworks for speaker verification antispoofing," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 4, pp. 684–694, Jun. 2017.

[13] J. Monteiro, J. Alam, and T. H. Falk, "Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers," *Comput. Speech Lang.*, vol. 63, 2020, Art. no. 101096.

[14] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, "Generalization of audio deepfake detection," in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, 2020, pp. 132–137.

[15] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 1068–1072.

[16] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Process. Lett.*, vol. 28, pp. 937–941, 2021.

[17] Z. Wu, R. K. Das, J. Yang, and H. Li, "Light convolutional neural network with feature genuinization for detection of synthetic speech attacks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 1101–1105.

[18] R. T. P., P. R. Aravind, R. C., U. Nechiyil, and N. Paramparambath, "Audio spoofing verification using deep convolutional neural networks by transfer learning," 2020, *arXiv:2008.03464*.

[19] H. Tak, J. Patino, A. Nautsch, N. Evans, and M. Todisco, "Spoofing attack detection using the non-linear fusion of sub-band classifiers," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 1106–1110.

[20] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 4259–4263.

[21] A. Luo, E. Li, Y. Liu, X. Kang, and Z. J. Wang, "A capsule network based approach for detection of audio spoofing attacks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 6359–6363.

[22] X. Li, X. Wu, H. Lu, X. Liu, and H. Meng, "Channelwise gated Res2Net: Towards robust detection of synthetic speech attacks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 4314–4318.

[23] X. Li *et al.*, "Replay and synthetic speech detection with Res2Net architecture," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 6354–6358.

[24] X. Ma, T. Liang, S. Zhang, S. Huang, and L. He, "Improved light CNN with attention modules for ASV spoofing detection," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.

[25] N. M. Muller, F. Dieckmann, P. Czempin, R. U. Canals, K. Bottinger, and J. Williams, "Speech is silver, silence is golden: What do ASVspoof-trained models really learn?," in *Proc. ASVspoof Workshop*, 2021, pp. 55–60.

[26] P. Khosla *et al.*, "Supervised contrastive learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 18661–18673.

[27] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[28] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[29] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[30] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[31] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2242–2251.

[32] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *Proc. Int. Conf. Learn. Representations*, 2019.

[33] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.

[34] A. van den Oord *et al.*, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.

[35] S. Huang, Q. Li, C. Anil, X. Bao, S. Oore, and R. B. Grosse, "TimbreTron: A WaveNet(CycleGAN(CQT(Audio))) pipeline for musical timbre transfer," in *Proc. Int. Conf. Learn. Representations*, 2019.

[36] N. Frosst, N. Papernot, and G. Hinton, "Analyzing and improving representations with the soft nearest neighbor loss," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2012–2020.

[37] X. Wang *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Comput. Speech Lang.*, vol. 64, pp. 101–114, 2020.

[38] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the Big Data paradigm with compact transformers," 2021, *arXiv:2104.05704*.

[39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, 2019, pp. 4171–4186.

[40] T. Kinnunen *et al.*, "t-DCF: A detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2195–2210, 2020.

[41] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007.