Dynamic Ensemble Teacher-Student Distillation Framework for Light-weight Fake Audio Detection

Jun Xue, *Student Member, IEEE,* Cunhang Fan, *Member, IEEE,* Jiangyan Yi, Jian Zhou, Zhao Lv, *Member, IEEE*

***Abstract*—In recent years, fake audio detection (FAD) has made great progress, and lightweight is important to achieve fast and reliable audio authenticity verification on resource-limited devices. However, most of the researchers ignore lightweight when improving the performance of FAD. To develop the application of FAD for small-end devices, this paper proposes a novel light-weight network named Light-ECA2Net. Given that networks with different depths have different abilities in capturing fake speech artifacts, this paper proposes a dynamic ensemble teacher-student distillation framework to fully transfer distillation knowledge. The dynamic ensemble distillation is divided into two aspects. First, we adopt one-to-one feature mapping to perceive the multidimensional feature knowledge and dynamically adjust every dimension feature weight by using ground truth labels, which can enable students to receive feature knowledge efficiently. Secondly, different network layers also have their strengths of predicting, further dynamically predicting weight can improve the learning ability of the student. Experimental results on the ASVspoof 2019 LA and PA datasets show that compared to the baseline, our system further improves performance by reducing the model complexity by 45%.**

***Index Terms*—Fake audio detection, Dynamic Ensemble distillation, ECA2Net**

I. Introduction

TEchnologies such as automatic speaker verification (ASV) and speech recognition are being used in various devices such as smart homes and smartphones. However, the rise of Deepfake technology brings unprecedented security challenges [1], [2], [3], [4], [5]. Fake audio detection (FAD) [6] can effectively prevent such deepfake attacks, and current research mainly utilizes the forging mechanism to build a robust detection system.

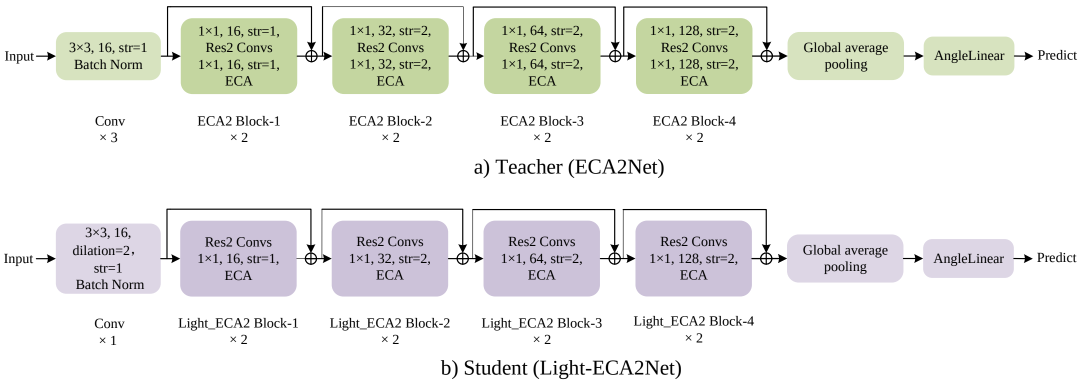
Researchers have developed a series of detection frameworks grounded in the mechanisms of synthetic speech. These frameworks can be broadly categorized into three types: 1) The first type employs neural networks based on ResNet and other convolutional neural networks to model manual acoustic features [7], [8], [9]. Examples include LFCC-LCNN [10] (LFCC, linear frequency cepstral coefficients), LPS-SENet [11] (LPS, log power spectrogram), and others. 2) The second type utilizes neural networks based on graph structures to extract discriminative information from original speech waveforms [12], [13]. Examples encompass GAT [14], and AASIST [15], among others. However, it's worth noting that modeling for graph nodes often demands more computational resources. 3) The third type leverages large-scale self-supervised language models, such as Wav2vec 2.0 [16], [17], HuBERT [18], etc., for fine-tuning. While these self-supervised pre-trained front-end features rely on extensive data, further fine-tuning necessitates a significantly high level of computational resources. In addition, some studies explore the complementarity between different detection models for fusion [3]. While this approach can improve performance, the multiple subsystems still concurrently escalate their complexity.

Knowledge distillation (KD) [19], [20], [21] emerges as a technique for transferring knowledge from one model to another, finding widespread use in tasks like model compression, acceleration, and accuracy optimization in deep learning. In the domain of FAD, researchers have also explored effectively in improving generalisability or lightweighting based on KD methods. For example, Xue et al. [22] proposed a universal self-distillation method to improve the performance of FAD models. In [21], the authors proposed a one-class knowledge distillation method based on Wav2Vec 2.0 to enhance the generalisation ability of the model. Further, FAD systems are usually deployed in edge devices, which imposes some limitations on computational resources and device space. For this reason, some research [23], [24], [25] has proposed a series of light-weight FAD systems through the KD method. However, these are difficult to balance the model size and performance well.

To further develop the lightweight application of FAD for small-end devices, this paper proposes a lightweight network (Light-ECA2Net) and a dynamic ensemble teacher-student distillation framework. In general, the deepest layer of a trained model is used for inference due to its high semantic features. Many studies [28], [29] on FAD have emphasized the importance of shallow knowledge, such as spectral defects and silent segments. Therefore, this paper effectively combines shallow and deep knowledge to prevent discriminative information from being lost during transmission. Specifically, we use the baseline ECA2Net as the teacher and the Light-ECA2Net as the student. Dynamic ensemble distillation (DED) framework can be divided into ensemble feature distillation

This work is supported by the STI 2030—Major Projects (No. 2021ZD0201500), the National Natural Science Foundation of China (NSFC) (No.62201002, 62322120), Distinguished Youth Foundation of Anhui Scientific Committee (No. 2208085J05), Special Fund for Key Program of Science and Technology of Anhui Province (No. 202203a07020008), Open Fund of Key Laboratory of Flight Techniques and Flight Safety, CACC (No, FZ2022KF15).). *(Corresponding authors: Cunhang Fan.)*

Jun Xue, Cunhang Fan, Jian Zhou and Zhao Lv are with the Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail:junxue.tech@gmail.com;cunhang.fan@ahu.edu.cn;Jzhou@ahu.edu.cn; kjlz@ahu.edu.cn) Jiangyan Yi is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 065001, China. (jiangyan.yi@nlpr.ia.ac.cn);



a) Teacher (ECA2Net)

b) Student (Light-ECA2Net)

Fig. 1. Framework and comparison of ECA2Net and Light-ECA2Net. ECA2Net is the embedding of ECA [26] block into the Res2Net [27] backbone. Light-ECA2Net is derived from further refinement of the ECA2Net architecture with a 45% reduction in the number of parameters.

(EFD) and ensemble logarithmic distillation (ELD). EFD employs one-to-one feature mapping to understand multi-dimension feature knowledge and dynamically adapts by using ground truth labels to receive feature knowledge efficiently from the teacher. According to the different network layers which also have their strengths, ELD dynamically ensembles all of the predictions to improve the learning ability of the student. Experimental results on ASVspoof 2019 LA and PA datasets show that our system reduces the model complexity by 45% while improving the performance compared to the baseline.

The remainder of this article is structured as follows: Section II outlines the proposed method. Section III presents the experimental setup, including the conducted experiments and the obtained results. Lastly, Section IV presents the conclusions.

II. The Proposed Method

We use ECA2Net as the teacher and Light-ECA2Net as the student. We propose a DED method to improve the performance of Light-ECA2Net. DED performs efficient knowledge transfer from two perspectives, feature transfer through one-to-one feature mapping and dynamic ensemble by using labels to compute the confidence of multidimensional features. The prediction dimension further enhances student learning ability by ensembling all shallow-deep prediction results.

A. The Teacher and Student Model

Fig. 1a shows the overall architecture and parameters of the ECA2Net, where Res2 Convs represents the intra-channel group residual structure. Fig. 1b shows our improved Light-ECA2Net model, which mainly uses a dilatation convolution to replace the three convolution operations on the input, and also reduces the ECA2 block internally.

B. Dynamic Ensemble Distillation

The proposed teacher-student framwork is shown as Fig. 2.

1) Ensemble logit distillation: First, we divide both the teacher model and the student model into four layers according to the network depth. After each layer, a classifier is added to obtain the prediction results of each layer, specifically expressed as:

$$P_i^t = \text{Teacher}\left(\text{Layer}_i\right); P_i^s \quad = \text{Student}\left(\text{Layer}_i\right) \qquad (1)$$

Where P_i^t and P_i^s represent the predictions of each layer of the teacher model and the student model, respectively.

To fully utilize the knowledge of networks, we integrate the teachers' knowledge by calculating the loss as the confidence level, which is expressed as:

$$\mathcal{L}_i^t = A\text{-softmax}\left(P_i^t, target\right); \mathcal{W}_i^t = \left(\mathcal{L}_i^t\right)^{-1} / \sum_{i=1}^{4}\left(\mathcal{L}_i^t\right)^{-1} \qquad (2)$$

where $\mathcal{W}_i^t$ is the ensemble weight.

The predictions of these four layers are also combined with the training of the student model, which can be expressed as follows:

$$\mathcal{L}_i^s = A\text{-softmax}\left(P_i^s, target\right); \mathcal{W}_i^s = \left(\mathcal{L}_i^s\right)^{-1} / \sum_{i=1}^{4}\left(\mathcal{L}_i^s\right)^{-1} \qquad (3)$$

Next, the ELD loss is calculated using the KL divergence function, which is described as:

$$\mathcal{L}_{ELD} = KL\left(\sum_{i=1}^{4} W_i^s * P_i^s, \sum_{i=1}^{4} W_i^t * P_i^t\right) \qquad (4)$$

2) Ensemble feature distillation: Each layer of the teacher and student models output corresponding features. The student learns the teacher's high-dimensional features through a multi-layer network under the fine guidance of EFD. Both the teacher and student models calculate feature distillation loss at each layer. The final EFD loss is obtained by weighting the confidence:

$$\mathcal{L}_{EFD} = \sum_{i=1}^{4} \mathcal{W}_i^t * \left\{MSE\left(\mathcal{F}_i^s, \mathcal{F}_i^t\right)\right\} \qquad (5)$$

In addition to the ELD and the EFD loss, the student needs to calculate the loss with the label during training.

$$\mathcal{L}_{hard} = A\text{-softmax}(output, target) \qquad (6)$$

This article has been accepted for publication in IEEE Signal Processing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LSP.2024.3431936

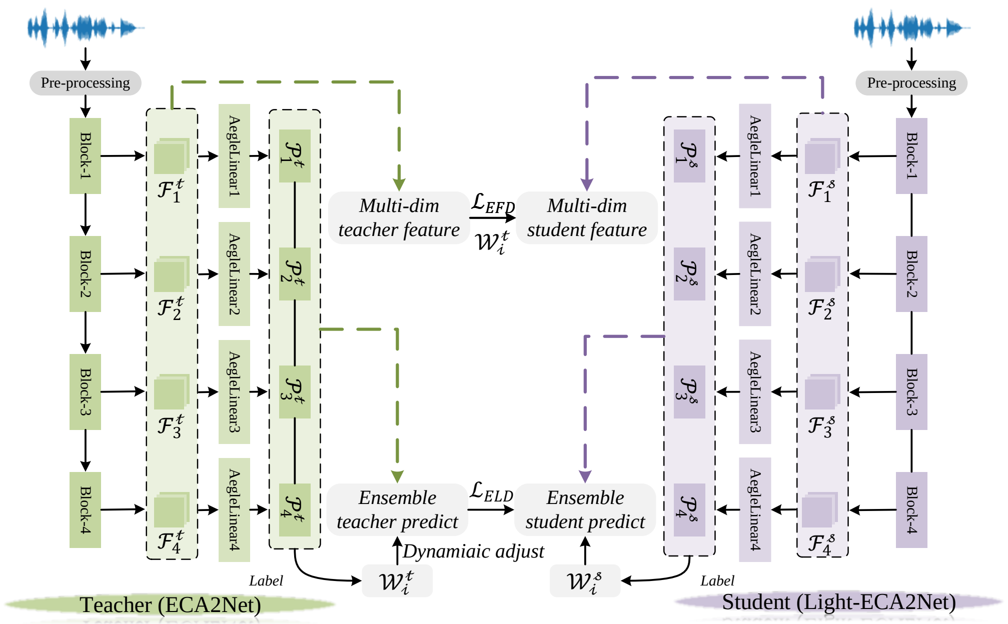


Fig. 2. The proposed dynamic ensemble teacher-student distillation framework. First, the teacher ensembles multi-dimensional features and dynamically adjusts the feature distillation loss using confidence weights. Further the teacher ensembles all shallow-deep network predictions and bounds the logit distillation loss by confidence weights. The teacher and student blocks correspond to the ECA2 Block and Light_ECA2 Block in Fig. 1, respectively. F_i^t and F_i^s denote the feature maps of each block, respectively. P_i^t and P_i^s denote the classification results of each block, respectively. W_i^t and W_i^s denote the dynamic distillation confidence of teachers and students, respectively.

$$\mathcal{L}_{ED} = \alpha * \mathcal{L}_{hard} + (1-\alpha) * \mathcal{L}_{ELD} + \beta * \mathcal{L}_{EFD} \qquad (7)$$

$\mathcal{L}_{ED}$ is the final training loss, and α and β are used to balance the three losses.

III. Experimental Results and Analysis

A. Dataset and Evaluation Metrics

The experiments in this paper are all based on the ASVspoof 2019[1] datasets. The ASVspoof 2019 LA dataset contains voice spoofing samples generated by three different speech synthesis technologies and six voice conversion technologies, as well as real voice samples. The ASVspoof 2019 PA dataset contains replay attacks, with a total of 27 types of replay attacks in different acoustic environments. Datasets are divided into a training set, a development set, and an evaluation set.

We use equal error rate (EER) and the minimum tandem detection cost function (t-DCF) as an evaluation metric. In addition, we use multiply-accumulates[2] (MACs) and parameters (param) as a measure of the model complexity and computational requirements of the neural network. MACs are the total number of multiply-accumulate operations performed in the model, and Param denotes the number of parameters in the model.

B. Feature Extraction and Experimental Settings

Following the literature [22], we chose the F0 subband as our input feature. We first apply the Blackman window function to perform a short-time Fourier transform to extract LPS, setting the window length and jump length to 1728 and 130, respectively. We set the frame length to 600, so that the extracted LPS feature dimension is 865 × 600. Finally, we cut out the first 45 dimensions in the frequency dimension to obtain the F0 subband, which has a feature dimension of 45 × 600. During the training process, we used Adam as the optimizer with parameters set to $\beta_1 = 0.9, \beta_2 = 0.98, \varepsilon = 10^{-9}$. The number of training rounds was set to 32. α and β are set to 0.7 and 0.3, respectively.

TABLE I
THE RESULTS OF THE ABLATION EXPERIMENT ON THE ASVSPOOF 2019 DATASET. LD AND FD STAND FOR LOGIT DISTILLATION AND FEATURE DISTILLATION USING ONLY THE LAST LAYER, RESPECTIVELY.

	Model	EER (%)	t-DCF	#Parm	MACs(G)
LA dataset	Teacher	0.76	0.2523	138K	23.33
	Student	1.56	0.0516	**76K**	**7.89**
	ED	**0.74**	**0.0246**	-	-
	ED - *w/o* ELD	0.98	0.0352	-	-
	ED - *w/o* EFD	0.95	0.0318	-	-
	Only LD	1.26	0.0430	-	-
	Only FD	1.11	0.0345	-	-
PA dataset	Teacher	0.75	0.0207	138K	23.33
	Student	0.77	0.0198	**76K**	**7.89**
	ED	**0.52**	**0.0146**	-	-
	ED - *w/o* ELD	0.63	0.0179	-	-
	ED - *w/o* EFD	0.59	0.0149	-	-
	Only LD	0.72	0.0194	-	-
	Only FD	0.67	0.0198	-	-

C. Experimental Results and Analysis

1) Experimental results based on ASVspoof 2019 LA dataset: In order to explore the impact of different distillation

[1] https://datashare.ed.ac.uk/handle/10283/3336

[2] https://github.com/Lyken17/pytorch-OpCounter

TABLE II
CAMPARISON OF PROPOSED METHOD WITH STATE-OF-THE-ART SINGLE SYSTEMS ON THE ASVSPOOF 2019 DATASET

Systems		Input Feature	EER(%)	t-DCF	#Param	MACs(G)
LA dataset	Light-ECA2Net (**ours**)	F0 subband	**0.74**	0.0246	**76K**	7.89
	Dual-Branch Network [30]	LFCC, CQT	0.80	**0.0210**		
	AASIST [15]	Waveform	0.83	0.0275	295K	219.03
	ECANet18_SD [22]	F0 subband	0.88	0.0295	711K	11.89
	AASIST-L [15]	Waveform	0.99	0.0309	85K	147.59
	SENet [11]	LPS	1.14	0.0368	1100K	32.95
	MCG-Res2Net50 [31]	CQT	1.78	0.0520	960K	108.5
	StuNet-OKD [24]	LFCC	2.23	0.0580	9510K	-
	StudentNet-FD [25]	LFCC	2.24	0.0580	409K	**0.29**
PA dataset	SE-Res2Net50 [28]	CQT	**0.52**	**0.0134**	920K	97.10
	Light-ECA2Net (**ours**)	F0 subband	0.53	0.0160	**76K**	7.89
	SENet34_SD [22]	F0 subband	0.65	0.0174	1344K	24.24
	T10 [32]	GD	1.08	0.1598	-	-
	T44 [32]	Log-DFT	1.29	0.1666	-	-
	StuNet-OKD [24]	LFCC	1.46	0.0390	9510K	-
	T53 [32]	log Mel grams	1.66	0.1729	-	-
	StudentNet-FD [25]	LFCC	1.69	0.0500	409K	**0.29**
	AASIST* [33]	Waveform	25.80	0.6850	295K	219.03

* This result is a replication of existing research [33].

strategies on the FAD system, this paper first trains the teacher and student model, and lets the teacher guide the student's training with different distillation strategies. The experimental results are shown in Table I. From Table I, we can see the following points: 1) Compared with the teacher model, the parameter volume of the student model is reduced by 45% and the computation is reduced by 66%, but as the parameters decrease, its EER also increased by 51%, indicating that the complexity of the model is closely related to the performance of FAD; 2) Comparing the student models under the guidance of different distillation strategies, it can be seen that our proposed ED method performs the best and even outperforms the teacher, which may be because the student model can perceive the feature details of the shallow network and learn the semantic representation of high-dimensional features under the guidance of the teacher model.

2) Experimental results based on ASVspoof 2019 PA dataset: To verify the generality of our proposed method, we also used the ASVspoof 2019 PA dataset. Table I shows the experimental results based on different distillation strategies. From Table I, we can see that compared to the baseline student model, its EER is 0.76%. The student models under different distillation strategies all show improvements, and both general logit distillation and feature distillation effectively improve the performance of the student model, indicating that the knowledge imparted by the teacher is very effective. Moreover, surprisingly, the student model guided by ensemble distillation can outperform the teacher, which is consistent with the results on the LA dataset, and the EER can even reach 0.53%, far exceeding the performance of the teacher model. This shows that under the ensemble distillation method, the student model perceives the shallow and deep knowledge of the teacher model, and also obtains discriminative feature information that cannot be obtained under normal training, which is very important for FAD. Further, we believe that ASVspoof 2019 PA has a more consistent distribution of training and test data. This allows the more lightweight model to both avoid overfitting when fitting the data and obtain more discriminative features through knowledge transfer from the teacher model, which may be the main reason why students outperform teachers in the PA scenario.

3) Comparison with other state-of-the-art single systems: To further validate the effectiveness of our proposed method, we compared our method with other recent advanced methods. Table II shows the results of other recent advanced methods on the ASVspoof 2019 LA and PA datasets. First, on the ASVspoof 2019 LA dataset, we can see that our method obtains the best result in terms of performance, model complexity, and computation. Compared with the current best model AASIST, which has an EER of 0.83%, parameters volume of 295K, and computation of 219.03G, our proposed model improves performance by 11% while reducing parameter volume by 74% and computation by 96%. On the ASVspoof 2019 PA dataset, Table 4 indicates our proposed single system that is essentially on par with the state-of-the-art in terms of performance, and has a greater advantage in terms of parameters and computation.

IV. CONCLUSION

In recent years, research on FAD has attracted much attention, but its development is still not mature enough compared to synthesized speech technology, and there are many problems. The problems addressed in this paper center around the convenience of FAD, which is important for the deployment of some mobile devices. In this paper, we first design ECA2Net and Light-ECA2Net as the baseline. To address the problem of degradation of the performance of the low parameters, we propose an DED method with ECA2Net as the teacher and Light-ECA2Net as the student. The proposed teacher-student framework improves the performance of the student model by aggregating shallow-deep logit and features, and the teacher guides the student from the perspective of multidimensional features and multilayered predictive knowledge, and the dynamic ensemble strategy not only enriches the distillation knowledge but also makes the distillation process more efficient. The experimental results show that the proposed system can achieve SOTA both in terms of parameters and performance. In the future, we should extend this lightweight study in complex scenarios.

References

[1] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, “ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge,” in *Proc. Interspeech*, 2015, pp. 2037–2041.

[2] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, “The ASVspoof 2017 challenge: assessing the limits of replay spoofing attack detection,” in *Proc. Interspeech*, 2017, pp. 2–6.

[3] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, “ASVspoof 2019: future horizons in spoofed and fake audio detection,” in *Proc. Interspeech*, 2019, pp. 1008–1012.

[4] 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.

[5] 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*. IEEE, 2022, pp. 9216–9220.

[6] M. Li, Y. Ahmadiadli, and X.-P. Zhang, “Audio anti-spoofing detection: A survey,” *arXiv preprint arXiv:2404.13914*, 2024.

[7] C. Fan, J. Xue, J. Tao, J. Yi, C. Wang, C. Zheng, and Z. Lv, “Spatial reconstructed local attention res2net with f0 subband for fake speech detection,” *Neural Networks*, p. 106320, 2024.

[8] C. Fan, J. Xue, S. Dong, M. Ding, J. Yi, J. Li, and Z. Lv, “Subband fusion of complex spectrogram for fake speech detection,” *Speech Communication*, vol. 155, p. 102988, 2023.

[9] R. Liu, J. Zhang, and G. Gao, “Multi-space channel representation learning for mono-to-binaural conversion based audio deepfake detection,” *Information Fusion*, vol. 105, p. 102257, 2024.

[10] X. Ma, T. Liang, S. Zhang, S. Huang, and L. He, “Improved lightcnn with attention modules for asv spoofing detection,” in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[11] Y. Zhang, W. Wang, and P. Zhang, “The effect of silence and dual-band fusion in anti-spoofing system,” in *Proc. Interspeech*, 2021, pp. 4279–4283.

[12] Y. Zhang, Z. Li, J. Lu, W. Wang, and P. Zhang, “Synthetic speech detection based on the temporal consistency of speaker features,” *IEEE Signal Processing Letters*, 2024.

[13] B. Huang, S. Cui, J. Huang, and X. Kang, “Discriminative frequency information learning for end-to-end speech anti-spoofing,” *IEEE Signal Processing Letters*, vol. 30, pp. 185–189, 2023.

[14] 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.

[15] 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*. IEEE, 2022, pp. 6367–6371.

[16] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, “Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation,” in *The Speaker and Language Recognition Workshop*, 2022.

[17] X. Wang and J. Yamagishi, “Can large-scale vocoded spoofed data improve speech spoofing countermeasure with a self-supervised front end?” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 311–10 315.

[18] J. Y. Xin Wang, “Investigating Self-Supervised Front Ends for Speech Spoofing Countermeasures,” in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 100–106.

[19] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *stat*, vol. 1050, p. 9, 2015.

[20] C. Wang, Y. Yue, B. Luo, Y. Chen, and J. Xue, “Psekd: Phase-shift encoded knowledge distillation for oriented object detection in remote sensing images,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 2680–2684.

[21] J. Lu, Y. Zhang, W. Wang, Z. Shang, and P. Zhang, “One-class knowledge distillation for spoofing speech detection,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 251–11 255.

[22] 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.

[23] Y.-L. Liao, X. Chen, C.-C. Wang, and J.-S. R. Jang, “Adversarial speaker distillation for countermeasure model on automatic speaker verification,” in *Proc. 2nd Symposium on Security and Privacy in Speech Communication*, 2022, pp. 30–34.

[24] Y. Ren, H. Peng, L. Li, X. Xue, Y. Lan, and Y. Yang, “A voice spoofing detection framework for iot systems with feature pyramid and online knowledge distillation,” *Journal of Systems Architecture*, vol. 143, p. 102981, 2023.

[25] Y. Ren, H. Peng, L. Li, and Y. Yang, “Lightweight voice spoofing detection using improved one-class learning and knowledge distillation,” *IEEE Transactions on Multimedia*, 2023.

[26] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, “Eca-net: Efficient channel attention for deep convolutional neural networks,” *CVPR 2020*, pp. 11 531–11 539, 2020.

[27] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. Torr, “Res2net: a new multi-scale backbone architecture,” *IEEE transactions on pattern analysis and machine intelligence*, pp. 652–662, 2019.

[28] X. Li, N. Li, C. Weng, X. Liu, D. Su, D. Yu, and H. Meng, “Replay and synthetic speech detection with res2net architecture,” in *ICASSP 2021*. IEEE, 2021, pp. 6354–6358.

[29] J. Deng, T. Mao, D. Yan, L. Dong, and M. Dong, “Detection of synthetic speech based on spectrum defects,” in *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, 2022, pp. 3–8.

[30] K. Ma, Y. Feng, B. Chen, and G. Zhao, “End-to-end dual-branch network towards synthetic speech detection,” *IEEE Signal Processing Letters*, vol. 30, pp. 359–363, 2023.

[31] X. Li, X. Wu, H. Lu, X. Liu, and H. Meng, “Channel-wise gated res2net: towards robust detection of synthetic speech attacks,” *Proc. Interspeech 2021*, 2021.

[32] 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.

[33] J. Kim and S. M. Ban, “Phase-aware spoof speech detection based on res2net with phase network,” in *ICASSP 2023*. IEEE, 2023, pp. 1–5.