MULTI-TASK LEARNING IMPROVES SYNTHETIC SPEECH DETECTION

Yichuan Mo, Shilin Wang

School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University
{mo536226,wsl}@sjtu.edu.cn

ABSTRACT

With the development of deep learning, synthetic speech has become more and more realistic and easier to spoof Automatic Speaker Verification (ASV) devices. Based on mining more effective hand-crafted features and proposing more powerful networks, many algorithms have been proposed to detect this malicious attack. In this paper, by observing that deepening the network impairs the performance of the network in detecting unknown attacks, we propose that the synthetic speech detection problem is an out-of-distribution (OOD) generalization problem and we enhance the robustness of networks by using multi-task learning. In our system, three auxiliary tasks are used to assist synthetic speech detection: bonafide speech reconstruction, spoofing voice conversion and speaker classification. Experimental results show that our approach can be applied to multiple architectures and can significantly improve the performance on both known attacks (development set) and unknown attacks (evaluation set). In addition, our best-performing network is quite competitive to recent state-of-the-art (SOTA) systems. It demonstrates the potential application of multi-task learning in synthetic speech detection.

Index Terms— synthetic speech detection, multi-task learning, speech anti-spoofing, adversarial learning

1. INTRODUCTION

While deep learning promotes the development of speech recognition, it also poses a lot of potential threats. One of those is synthetic speech which is widely discussed in the community. Divided into Voice Conversion (VC) [1, 2] and Text to Speech (TTS) [3, 4], high-quality synthetic speech is now able to fool both humans and voice recognition systems easily. In order to drive more attention to this problem, ASVspoof challenge [5, 6, 7] has been held for years and a lot of countermeasures have been developed to defend against these attacks. On the one hand, some previous works mainly focused on extracting more representative features: To reflect more defects of synthetic speech, linear frequency cepstral coefficients (LFCCs) were used in the local binary pattern proposed in [8]. Utilizing the hardness of modeling the dynamic behavior of bonafide speech, the authors in [9] found that traditional dynamic coefficients are helpful for synthetic speech detection , so dynamic LFCC was proposed. Motivated by constant Q transform, [10] reported a new feature, constant Q cepstral coefficients (CQCCs) can bring greater robustness to detecting unforeseen spoofing attacks. On the other hand, many efforts have been made to improve network structures: combining Gated Recurrent Unit (GRU) and Sinc filter, Rawnet2 [11] has the potential to capture artifacts of unknown attacks. Graph attention networks (GATs) in [12] uses a self-attention mechanism to model the relationships between neighboring sub-bands which is the drawback of the synthetic speech. To enable multiple feature scales, Res2Net [13] modifies the ResNet[14] block to improve the generalizability of detectors. [15] redesigned the original capsule network to make it suitable for the forensic task. Unlike most of the mainstream approaches, we first report an intriguing phenomenon: increasing the depth of the neural network will decrease the robustness of the network against unknown attacks. Motivated by this property, we propose to enhance synthetic speech detection by sharing representations with other tasks, i.e., multi-task learning. Shown in Fig. 1, our system is assisted by three subtasks: reconstructing the real speech using an inverse network, generating more realistic samples by applying adversarial learning, classifying the embedding features of bonafide speech. Experimental results show that our proposed method can be applied to multiple frameworks and can improve the capability of the detector for detecting malicious synthetic speech. In short, our contributions are summarized as follows:

- By showing the degradation problem of ResNet, we illustrate that synthetic speech detection is an out-of-distribution (OOD) generalization problem.
- We propose a three-task-assisted training architecture to enable the detector to learn cross-domain features, which can be simultaneously integrated with different network structures and input features.
- Results on the benchmark dataset show the notable improvement of our approach on seen and unseen attacks, whereby the performance gain is orthogonal to those from existing synthetic speech detection methods.

This work was supported by the National Natural Science Foundation of China (61771310). Shilin Wang is the corresponding author.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

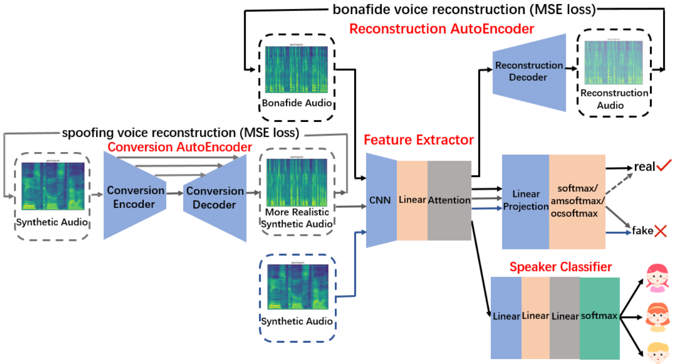
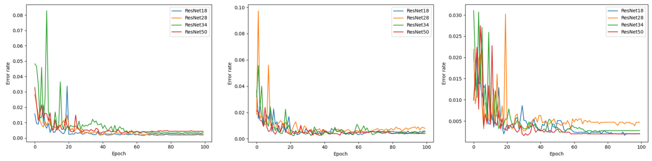


Fig. 1: Structure of our synthetic speech detection network. Each color of lines represents the flow of one kind of features. When training Conversion AutoEncoder, we maximize the likelihood that the reconstructed feature is classified as natural speech. However, when we train the detector, we classify those features as attacks.

2. BACKGROUND

2.1. Overfitting in specific domains

Utilizing the open-source code[1] from Github, we train the networks of different depths under the same hyperparameter configuration (long enough for each network to converge, Fig.2). Consistent with Asvspoof2019 competition, we take tandem detection cost function (t-DCF) [16] and equal error rate (EER) as the evaluation metrics. One thing to note is that the structures of ResNet28 and ResNet34 are the same in the origin code, so we fix this bug by setting $[3; 4; 4; 2]$ PreActBlocks as the configuration for ResNet28. We perform experiments on three kinds of loss functions: softmax, am-softmax [17] and oc-softmax [18]. To avoid overfitting on the training set, we test the performance of models on the evaluation set using the checkpoint with the lowest error rate on the development set. The results in Table 1 reveal that for almost all loss functions, deepening the network is detrimental to the network's ability to generalize on unknown attacks. It illustrates the big difference between ordinary classification (e.g. CIFAR-10, ImageNet) and synthetic speech detection where the detector overfits on the domain-specific features. Therefore, we can no longer treat the different synthetic speech generated by different synthesis methods as independent, identical distribution (IID) but rather as out-of-distribution (OOD) data. Discussed extensively in visual tasks, natural factors are a large contributor to distribution shifts [19], but for synthetic speech, the synthesis method can have the main impact on the synthetic speech features. Therefore, we hope to detect synthesized speech based on features shared by multiple synthesis methods rather than domain-specific features and our approach focuses on learning cross-domain features using multi-task learning.



(a) softmax (b) am-softmax (c) oc-softmax

Fig. 2: EER in developed partition of ASVspoof2019 dataset.

Table 1: Performance in evaluation set.

Model		ResNet18	ResNet28	ResNet34	ResNet50
softmax	EER(%)	**4.64**	5.23	5.68	5.00
	t-DCF	**0.109**	0.125	0.164	0.125
am-softmax	EER(%)	**2.56**	2.62	3.17	4.08
	t-DCF	0.065	**0.064**	0.066	0.095
oc-softmax	EER(%)	2.17	**2.15**	2.42	4.65
	t-DCF	**0.053**	0.057	0.060	0.136

2.2. Multi-task learning and anti-spoofing detection

First proposed by Caruana in 1997 [20], multi-task learning was defined as *"Learn tasks in parallel and share the representation to enable the original tasks to learn better."* Since then, it has been used in computer vision[21] and natural language processing[22]. More related to our work, [23] and [24] separately proposed auxiliary noise classifier and adversarial speaker recognition to better detect replayed speech. In contrast, our work focuses on synthetic speech detection, which has been little explored in previous works. Our system consists of three subtasks: reconstruct the natural speech in order to reduce the loss of human features in the feature extraction process, learn a tighter decision boundary via adversarial learning, classify features of real speech to obtain identity information before speech conversion. In Section 4, the networks which are assisted by auxiliary tasks can obtain huge improvements in both development and evaluation sets. It indicates that cross-domain features are indeed more beneficial for synthetic speech detection.

3. THE PROPOSED APPROACH

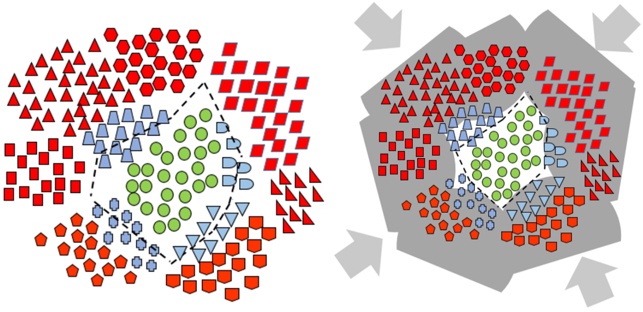
3.1. Bonafide speech reconstruction branch

The intuition behind this method is that by reconstructing the bonafide speech, we can extract more representative features of bonafide speech. This allows our detector to discriminate more based on the characteristics of natural persons rather than the defects of synthetic speech. [25] demonstrates that Autoencoders are able to capture the noisy input, so they reconstruct the CQCC features of replayed speech before sending them into the siamese network. Different from them, we use the reconstruction decoder D_{rec} to maximize a posterior, $P(\mathbf{x}|\mathbf{z})$, where $\mathbf{z}$ is the embedding feature of speech $\mathbf{x}$, extracted by feature extractor, F. For a minibatch of real samples: $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$, the reconstruction loss is defined as:

$$loss_{rec} = \frac{1}{n}\sum_{i=1}^{n} \left|\left|\left(D_{rec}(F(\mathbf{x}_i)) - \mathbf{x}_i\right)^2\right|\right|_1 \qquad (1)$$

When we train this branch, F and D_{rec} are jointly optimized to minimize $loss_{rec}$. Moreover, a trade-off coefficient λ_r is introduced to balance the main task and this subtask.

[1] https://github.com/yzyouzhang/AIR-ASVspoof



(a) Loose decision boundary.

(b) Tight decision boundary.

Fig. 3: Attacks in the training set, enhanced samples of the training set, unseen attacks, and natural speech are represented by four colors: red, grey, blue and green, respectively. Moreover, the black dotted line represents the decision boundary and each shape, except for the circle, represents a method of speech synthesis.

3.2. Spoofing voice conversion branch

Motivated by the adversarial learning in GANs [26], we use the zero-sum game between synthetic speech detector and Conversion Autoencoder (CA) to tighten the decision boundary: CA tries to deceive the detector by generating more realistic spoofing speech. However, as a defender, the detector tries to classify the converted speech as an attack. Unlike GANs, we aim to train a sharper discriminator by augmenting the original spoofing samples from the non-convergent nature of adversarial learning[27]. To avoid directly mapping to the real distribution, the reconstruction constraint is added to the converted samples. Apart from that, a coefficient, δ is adopted to avoid the over converging of CA (If δ is small enough, CA will converge to an "identity mapping" instead of a sample enhancer.). To better demonstrate the role of spoofing voice conversion, Fig. 3 is shown for illustration. Because of overfitting on domain-specific features, the decision boundary of the detector is relatively loose, thus limiting the generalization capability for defending unseen attacks. Through the dynamic interaction in adversarial learning, the spoofing samples are enhanced according to the weakness of the detector. It forces the detector to use more robust features to distinguish the natural and spoofing speech and learn more robust decision boundary. In order to better deliver high-dimensional information and reduce the difficulty of reconstruction, we use U-net [28] as our CA. For a minibatch of synthetic speech: $\{\mathbf{s}_1, ..., \mathbf{s}_n\}$, CA is trained with the following loss function:

$$loss_{at} = loss_{atr}(\mathbf{s_i}) + \delta loss_{ata}(\mathbf{s_i}) \tag{2}$$

$loss_{atr}$ and $loss_{ata}$ are given by the following equations:

$$loss_{atr} = \frac{1}{n}\sum_{i=1}^{n} \left|\left|\left(D_{con}(E_{con}(\mathbf{s}_i)) - \mathbf{s}_i\right)^2\right|\right|_1 \tag{3}$$

and

$$loss_{ata} = \frac{1}{n}\sum_{i=1}^{n} log\big(1 - C_{det}\big(F(D_{con}(E_{con}(\mathbf{s}_i)))\big)\big) \tag{4}$$

In Eq.3 and Eq.4, D_{con} and E_{con} stand for decoder or encoder used in voice conversion respectively. C_{det} denotes the classifier used to detect synthetic speech, consisting of two linear layers. For the synthetic classifier and feature extractor, their loss for discrimination is:

$$loss_{con} = \frac{1}{n}\sum_{i=1}^{n} log\Big(C_{det}(F(D_{con}(E_{con}(\mathbf{s}_i))))\Big) \tag{5}$$

Same as bonafide speech reconstruction, we also introduce λ_c as a coefficient to ensure synthetic speech detection is mainly considered when updating parameters.

3.3. Speaker classification branch

Gajan et al. proposed to train speaker recognition network and replay-detection network adversarially to acquire identity immutability [24]. It is reasonable because the drawbacks of synthetic speech is highly related to the recording method rather than identity for replay attacks. However, different from replay attacks, speech synthesis systems create the features instead of copying them, which causes variability in the quality of different features. The above difference can be used for synthetic speech detection and speaker classification simultaneously and is invariant to various TTS attacks. By sharing representation with this subtask, we hope to extract features that are widely available and difficult to generate for TTS methods instead of the characteristics of one or two persons, such as the speech speed. Thus it can alleviate the overfitting of the detector on the specific domains. In addition, by classifying the bonafide speech, we want to train a classifier that can extract the residual identity information before voice conversion. It can be applied to all synthesis speech generated by almost all speech conversion methods. In short, the discrepancy between different features and the similarities among VC attacks ensure speaker classification can assist the detection task. For a minibatch of real speech: $\{(\mathbf{x}_1, y_1)...(\mathbf{x}_2, y_2)\}$, the speaker classification loss is:

$$loss_{cls} = \frac{1}{n}\sum_{i=1}^{n} CE\big(C_{cls}(F(\mathbf{x}_i)), y_i\big) \tag{6}$$

Here CE is short for cross-entropy loss function. In addition, C_{cls} means the speaker classifier. λ_m is also used to balance the main task and subtask.

3.4. Loss function design

The loss for detecting synthetic speech is defined as:

$$loss_{de} = \frac{1}{n}\sum_{i=1}^{n} BCE\big(C_{det}(F(\mathbf{m}_i)), k_i\big) \tag{7}$$

Here BCE means binary cross-entropy loss function. For oc-softmax or am-softmax, its form would be changed accordingly. $\{\mathbf{m}_1, ...\mathbf{m}_n\}$ represents the total dataset, consisting of all real samples $\{\mathbf{x}_1, ...\mathbf{x}_n\}$ and all synthetic samples $\{\mathbf{s}_1, ...\mathbf{s}_n\}$. k_i is its corresponding ground-truth label about whether it is an attack. Combine the main task and all subtasks, the form of the total loss function is:

$$loss_{all} = loss_{de} + \lambda_m loss_{cls} + \lambda_c loss_{con} + \lambda_r loss_{rec} \tag{8}$$

Except for Conversion Autoencoder is updated using Eq. 2, other networks are jointly trained using the total loss function (Eq.8).

4. EXPERIMENT

4.1. Dataset and baseline system

We choose the logical access part of ASVspoof 2019 as our benchmark dataset. The training and development sets contain the same six types of attacks (two VC and four TTS). And other 13 unseen attacks are included in the evaluation set. To better demonstrate the positive impact of multi-task learning on synthetic speech detection, we conduct ablation experiments for softmax, am-softmax and oc-softmax, which are specifically designed for synthetic speech detection and the experimental results in [18] are chosen as our baseline (shown in Table 4). In addition, the results in Table 5 for other recent best-performing synthetic speech detection approaches are obtained from the original papers. All experiments below (including Table 1) are performed on single or multiple GeForce RTX 2080Ti GPU.

4.2. Multi-task learning for different structures

In this section, we test multi-task learning on multiple networks to illustrate the mitigation effect of our method on overfitting in specific domains. All three auxiliary branches are used and we select oc-softmax as our loss function. Because ResNet18 suffers from the underfitting problem, we add another network: ResNet40, composed of [4;5;7;3] PreActBlocks to enhance the persuasiveness of the experiment. All networks are trained with the same hyperparameter and the results are shown in Table 2. Regardless of the EER or the t-DCF metrics, our algorithm boosts the robustness of all architectures for detecting unknown attacks.

Table 2: Multi-task learning in multiple architectures

Model		ResNet28	ResNet34	ResNet40	ResNet50
EER(%)	baseline	2.15	2.42	2.46	4.65
	multi-task	1.79	1.47	2.09	3.05
t-DCF	baseline	0.057	0.060	0.059	0.136
	multi-task	0.046	0.036	0.055	0.075

4.3. Ablation studies

In our ablation studies, we choose the ResNet34 architecture as our backbone to test the effects of three subtasks separately and in combination on synthetic speech detection. S_1, S_2, S_3 respectively denote bonafide speech reconstruction, spoofing voice conversion and speaker classification. Moreover, the training strategy is consistent with the baseline networks. The performance of our systems on the development and evaluation partition is shown in Table 3 and Table 4. For the best multi-task systems, we achieve 33.7%, 36.2%, 32.9% improvement on softmax, am-softmax, oc-softmax structured networks of evaluation partition. One thing worth noting is that enhancing generalization in out-of-distribution domains is not at the expense of the performance in training domains: most systems achieve a significant improvement in the EER and t-DCF for both development and evaluation set. It proves that with the help of multi-task learning, synthetic speech detectors can learn cross-domain features that are more robust to all kinds of attacks.

Table 3: Ablation studies in development dataset

Model		Baseline	S_1	S_2	S_3	S_1+S_2	S_1+S_2+S_3
softmax	EER(%)	0.35	0.08	0.39	0.24	0.16	0.13
	t-DCF	0.010	0.002	0.010	0.004	0.004	0.003
am-softmax	EER(%)	0.43	0.43	0.39	0.27	0.36	0.36
	t-DCF	0.013	0.014	0.012	0.008	0.011	0.010
oc-softmax	EER(%)	0.20	0.20	0.24	0.19	0.16	0.16
	t-DCF	0.006	0.007	0.006	0.006	0.005	0.005

Table 4: Ablation studies in evaluation dataset

Model		Baseline	S_1	S_2	S_3	S_1+S_2	S_1+S_2+S_3
softmax	EER(%)	4.69	3.86	3.76	3.52	3.11	3.61
	t-DCF	0.125	0.085	0.091	0.079	0.077	0.089
am-softmax	EER(%)	3.26	2.29	2.15	2.08	2.08	2.32
	t-DCF	0.082	0.050	0.059	0.054	0.055	0.057
oc-softmax	EER(%)	2.19	1.82	2.12	2.11	1.70	1.47
	t-DCF	0.059	0.047	0.054	0.054	0.042	0.036

4.4. Comparison with the SOTA systems

Table 5: Comparison of our multi-task system to some known SOTA single systems

Systems	Front-end	t-DCF	EER(%)
RawGAT-ST[29]	Raw-audio	0.034	1.06
ResNet34+OC-softmax+S_1+S_2+S_3	**LFCC**	**0.036**	**1.47**
GMM+GAT-S+GAT-T+RawNet2 [12]	LFCC	0.048	1.68
ResNet34+OC-softmax+S_1+S_2	**LFCC**	**0.042**	**1.70**
CQT+MCG-Res2Net50+CE[30]	CQT	0.052	1.78
ResNet18-L-FM[31]	LFBs	0.052	1.81
ResNet34+OC-softmax+S_1	**LFCC**	**0.047**	**1.82**
LCNN-LSTM-sum[32]	LFCC	0.052	1.92
Capsule network[15]	LFCC	0.054	1.97
OC-softmax[18]	LFCC	0.059	2.19
GMM-fusion[33]	LFCC	0.074	2.92
Siamese CNN[34]	LFCC	0.093	3.79
GKDE-Triplet[35]	STFT	0.086	3.84
FG-LCNN[36]	CQT	0.102	4.07
FFT-LCNN[37]	FFT	0.103	4.53

In this section, we compare our multi-task system with known SOTA networks. The results in Table 5 illustrate that multi-task learning can enable ordinary ResNet34 to surpass most well-designed networks. As far as we know, we are the SOTA of the single systems for taking LFCC features as input, which shows the great potential of multi-task learning in synthetic speech detection. In addition, the proposed multi-task framework can be integrated with most model architectures, e.g. Rawnet [12], Res2Net[30], to obtain better performance for defending attacks from various malicious speech.

5. CONCLUSION

Based on the phenomenon that the increase in depth will reduce the capability for detecting unseen attacks, we illustrate that synthetic speech detection is an OOD problem. Therefore, our strategy is to learn cross-domain features by using multi-task learning. In this paper, our multi-task framework is composed of three branches: bonafide speech reconstruction, spoofing voice conversion and speaker classification. The effectiveness of our method is fully discussed and we hope that more works will be done in the future to propose better multi-task learning architectures to improve the performance of synthetic speech detection.

6. REFERENCES

[1] Takuhiro Kaneko and Hirokazu Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2100–2104.

[2] Rafael Ferro, Nicolas Obin, and Axel Roebel, "Softgan: Learning generative models efficiently with application to cyclegan voice conversion," *CoRR*, 2019.

[3] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[4] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[5] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Sixteenth annual conference of the international speech communication association*, 2015.

[6] Tomi Kinnunen, Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," 2017.

[7] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," in *INTERSPEECH 2019-20th Annual Conference of the International Speech Communication Association*, 2019.

[8] Timo Ojala, Matti Pietikainen, and Topi Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[9] Md Sahidullah, Tomi Kinnunen, and Cemal Hanilçi, "A comparison of features for synthetic speech detection," 2015.

[10] Massimiliano Todisco, Héctor Delgado, and Nicholas WD Evans, "A new feature for automatic speaker verification anti-spoofing: constant q cepstral coefficients.," in *Odyssey*, 2016, vol. 2016, pp. 283–290.

[11] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony 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.

[12] Hemlata Tak, Jee-weon Jung, Jose Patino, Massimiliano Todisco, and Nicholas Evans, "Graph attention networks for anti-spoofing," *arXiv preprint arXiv:2104.03654*, 2021.

[13] Xu Li, Na Li, Chao Weng, Xunying Liu, Dan Su, Dong Yu, and Helen Meng, "Replay and synthetic speech detection with res2net architecture," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6354–6358.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] Anwei Luo, Enlei Li, Yongliang Liu, Xiangui Kang, and Z Jane Wang, "A capsule network based approach for detection of audio spoofing attacks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6359–6363.

[16] Tomi Kinnunen, Kong Aik Lee, Héctor Delgado, Nicholas Evans, Massimiliano Todisco, Md Sahidullah, Junichi Yamagishi, and Douglas Reynolds, "t-dcf: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in *Speaker Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018.

[17] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[18] You Zhang, Fei Jiang, and Zhiyao Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.

[19] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al., "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8340–8349.

[20] Rich Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[21] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang, "Facial landmark detection by deep multi-task learning," in *European conference on computer vision*. Springer, 2014, pp. 94–108.

[22] Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang, "Multi-task learning for multiple language translation," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 1723–1732.

[23] Hye-Jin Shim, Jee-Weon Jung, Hee-Soo Heo, Sung-Hyun Yoon, and Ha-Jin Yu, "Replay spoofing detection system for automatic speaker verification using multi-task learning of noise classes," in *2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*. IEEE, 2018, pp. 172–176.

[24] Gajan Suthokumar, Vidhyasaharan Sethu, Kaavya Sriskandaraja, and Eliathamby Ambikairajah, "Adversarial multi-task learning for speaker normalization in replay detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6609–6613.

[25] Mohammad Adiban, Hossein Sameti, and Saeedreza Shehnepoor, "Replay spoofing countermeasure using autoencoder and siamese networks on asvspoof 2019 challenge," *Computer Speech & Language*, vol. 64, pp. 101105, 2020.

[26] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[27] Weili Nie and Ankit B Patel, "Towards a better understanding and regularization of gan training dynamics," in *Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 281–291.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[29] Hemlata Tak, Jee-weon Jung, Jose Patino, Madhu Kamble, Massimiliano Todisco, and Nicholas Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," *arXiv preprint arXiv:2107.12710*, 2021.

[30] Xu Li, Xixin Wu, Hui Lu, Xunying Liu, and Helen Meng, "Channel-wise gated res2net: Towards robust detection of synthetic speech attacks," *arXiv preprint arXiv:2107.08803*, 2021.

[31] Tianxiang Chen, Avrosh Kumar, Parav Nagarsheth, Ganesh Sivaraman, and Elie Khoury, "Generalization of audio deepfake detection," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 132–137.

[32] Xin Wang and Junich Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," *arXiv preprint arXiv:2103.11326*, 2021.

[33] Hemlata Tak, Jose Patino, Andreas Nautsch, Nicholas Evans, and Massimiliano Todisco, "Spoofing attack detection using the non-linear fusion of sub-band classifiers," *arXiv preprint arXiv:2005.10393*, 2020.

[34] Zhenchun Lei, Yingen Yang, Changhong Liu, and Jihua Ye, "Siamese convolutional neural network using gaussian probability feature for spoofing speech detection.," in *INTERSPEECH*, 2020, pp. 1116–1120.

[35] Alejandro Gomez-Alanis, Jose A Gonzalez-Lopez, and Antonio M Peinado, "A kernel density estimation based loss function and its application to asv-spoofing detection," *IEEE Access*, vol. 8, pp. 108530–108543, 2020.

[36] Zhenzong Wu, Rohan Kumar Das, Jichen Yang, and Haizhou Li, "Light convolutional neural network with feature genuinization for detection of synthetic speech attacks," *arXiv preprint arXiv:2009.09637*, 2020.

[37] G Lavrentyeva, A Tseren, M Volkova, A Gorlanov, A Kozlov, and S Novoselov, "Stc antispoofing systems for the asvspoof2019 challenge," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 1033–1037.