INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Learning A Self-Supervised Domain-Invariant Feature Representation for Generalized Audio Deepfake Detection

Yuankun Xie[1], *Haonan Cheng*[2], *Yutian Wang*[3], *Long Ye*[4]

[1]State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China

xieyuankun@cuc.edu.cn, haonancheng@cuc.edu.cn, wangyutian@cuc.edu.cn, yelong@cuc.edu.cn

Abstract

In recent years, Audio Deepfake Detection (ADD) models have shown promising results in intra-domain. However, they do not perform well in cross-domain scenarios. This is mainly due to the limited variety of domain types and attack methods in training data, as well as insufficient research on hidden feature representation. To address these issues, we present W2V-ASDG, a generalized ADD system including a self-supervised representation front-end and a domain generalization backbone. Furthermore, we try to learn an ideal feature space which aggregates real speech and separates fake speech. Fake speech varies significantly by different forgery methods, while real speech varies less. In light of this, we further propose the aggregation and separation domain generalization (ASDG) method as the back-end to learn a domain invariant feature representation. Experiments show that our W2V-ASDG outperforms baseline models in cross-domains and gets the lowest average equal error rates (EER) of 4.60%.

Index Terms: audio deepfake detection, self-supervised representation, domain generalization, feature space

1. Introduction

With the rapid development of text-to-speech and voice conversion, the adversarial technique of Audio Deepfake Detection (ADD) has attracted increasing interest. In current ADD methods, different strategies such as manual features [1] and Data Augmentation (DA) [2, 3] are frequently adopted, which show promising results in intra-domain database testing [4, 5]. Nonetheless, these methods deteriorate in performance when applied to cross-domain conditions [6]. This is primarily caused by the following two reasons:

- The complexity of the target speech (codec, channel, noise, etc.,) [7] and variety of the attack types may lead to misjudgment of ADD model.
- Recent ADD methods lack of generalization and interpretability of hidden representation. The existing methods rely on deep backbone networks such as Resnet [8] and LCNN [9], which may cause overfitting to the source domain.

To tackle the issue of performance degradation in cross-domain, we first take advantage of wav2vec2 (W2V2), a self-supervised model as front-end. W2V2 is trained on a large amount of real utterances from different source domains without any labels. The utilization of rich source domain information enhances our ability to discriminate in complex cross-domain scenarios (different codec, channel, noise, etc.,). Thus, the attributes of hidden states and variations of value between real and fake speech can help to set a discriminative class boundary. Recent studies have been investigated the effect of self-supervised front-end [10] and perform well on both intra-domains [11–13] and cross-domain [14, 15]. In our model, we utilize W2V2-XLS-R [16] as the front-end to extract the feature from raw audio as it is the largest version currently.

As we known, DA is a common strategy in most ADD tasks, which is used to adapt the test speech scene by augmenting the speech environment. However, in real-world scenarios, there are various situations to consider. It's inefficient to add new bonafide or spoofed audio to the training data to achieve better result. To overcome the challenge of generalizing to an unseen target domain with limited source data, Domain-Invariant Representation Learning (DIRL) techniques have been proposed. The goal of DIRL is to reduce the representation discrepancy between multiple source domains among various source domains, ensuring domain invariance. In recent years, DIRL strategies have been widely proven to effectively combat the decline in anti-spoofing generalization performance caused by illumination and camera changes [17–19]. Similar to face forgery data, speech also has very large differences due to different environmental noise, recording equipment, and deception attack methods. Therefore, It may be beneficial to introduce DIRL strategies into the ADD task.

Based on the aforementioned analysis, we propose W2V-ASDG, an ADD method with higher generalization ability based on DIRL. Initially, we hypothesize that in an ideal discriminative feature space, the data distribution of real speeches should be clustered into a single cluster regardless of domains, while the fake one should be more scattered. This is due to the fact that although various devices or channels may have some impact on both genuine and spoofing speech, the divergent attack types have a greater impact on the variation of magnitude for spoofing speech. Then, to construct the feature space, we aggregate the real speech distribution through a single-side adversarial domain discriminator and separate the fake speech by triplet mining. The input of the domain discriminator includes only real datasets rather than fake ones to make the features of real speech from different domains undistinguishable. Simultaneously, the triplet mining technique can effectively discrete the spoofing features and aggregate the real features.

The main contributions of this work are as follows:

- We propose a well-generalized ADD model, W2V-ASDG, which comprises a large-scale self-supervised front-end W2V2-XLS-R, and a domain generalization backbone ASDG.
- We introduce an ideal domain-invarient feature space recognition of the real and fake speech feature distribution. ASDG method is designed to achieve that goal, which improves interpretability.
- Numerous experiments are organized in harsh conditions:

10.21437/Interspeech.2023-1383

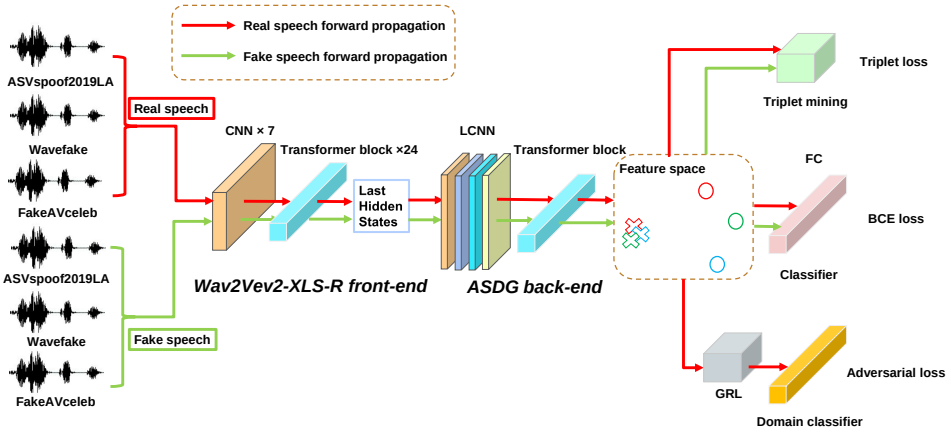

Figure 1: *The whole architecture of our proposed W2V-ASDG which include W2V2-XLS-R front-end and ASDG back-end. The red and green lines represent the forward propagation routes of real and fake speech respectively.*

cross-language and noisy dataset. Experiments show that our model outperforms baseline models in cross-domains and get the lowest EER for 4.60%, which can verify the generalization ability of our model on unknown target domains.

2. Proposed Method

The framework of our proposed system for ADD is depicted in Figure 1. Three different domains are used to enhance the diversity of training data. At first, W2V2-XLS-R is utilized to extract the feature of raw audio. Then, LCNN and transformer block as a backbone to make a binary classification decision. To complete an ideal feature space, a single-side adversarial domain discriminator is proposed to identify which domain the feature belongs to. In this way, we can aggregate the real speech regardless of which domain the speech belongs to. Furthermore, to separate the fake feature while aggregating the real one, the triplet mining method is added to achieve the goal. Finally, the overall loss is integrated to optimize the network.

2.1. W2V2-XLS-R front-end

In the left-side of Figure 1, W2V2 based front-end is trained by solving a contrastive task over a masked feature encoder. All of the speech signals from three domains are first processed by the feature extractor which is composed of seven convolutional neural network (CNN) layers. Then, the Transformer network which is composed by 24 layers, 16 attention heads, and 1024 embedding size are used to obtain context representations.

In the training phase, the feature extractor representations are discretized to a quantized vector to represent the targets in the objective. In practice, we employ the above steps by Hugging face version of wav2vec2-XLS-R-300M[1] and freeze the weights of front-end. The front-end model is pre-trained with 436k hours of unannotated genuine speech data in 128 languages using a contrastive objective. As a result of pre-training, the last hidden states from the transformer can represent the genuine speech contextualized information.

[1]https://huggingface.co/facebook/wav2vec2-xls-r-300m

2.2. ASDG back-end

2.2.1. Backbone

In order to obtain an ideal feature space from a limited dataset, a well-performing feature generator is essential. In feature generator, we use the light convolution network (LCNN) and transformer block to extract the feature space from the source domain speech. The details of LCNN structure is same as [9]. To capture the relations between adjacent elements on time series, a single SpeechFormer block are added after LCNN. The difference between SpeechFormer block and standard transformer is that the multi-head attention utilizes a window to limit the computation to a small scope of adjacent tokens, which can greatly relieve the computational burden [20].

2.2.2. Domain classifier

In an ideal feature space, the distribution of real speech is aggregated regardless of domain. Thus, a single-side adversarial domain discriminator with Gradient Reverse Layer (GRL) [21] is proposed. As we can see in Figure 1, the domain discriminator only discriminates the feature from real domains. Let $p(X_r)$ denotes the distributions of real feature and Y_D denotes the domain of X_r. The adversarial loss function of the domain discriminator is defined as follows:

$$\min_{D}\max_{G} L_{ada}(G, D) = - E_{x\sim P(X_r), y\sim Y_D} \sum_{d=1}^{3} p(y = d) log\left(D\left(G(x)\right)\right) \quad (1)$$

where d denotes the domain label. Minimizing loss in discriminator and maximizing loss in the generator are simultaneous. The feature generator is trained to learn a robustness feature to spoof the domain discriminator in order to maximize L_{ada}. In the meantime, the discriminator is trained to identify the feature domain by minimizing. To achieve the synchronous goal, GRL is added to make the discriminator unable to identify which domain the real feature originates from. GRL doesn't work in the forward propagation. Instead, GRL layer reverses the gradient by multiplying negative dynamic coefficients in the backward propagation to make the discrimination task difficult. As the number of training iter increases, the coefficient decreases from

0 to -1 so that the model will focus on optimizing the classifier loss at the beginning of the training process.

2.2.3. Triplet mining

Owing to the diversity of the spoofing method in ADD, the distribution of fake speech should be dispersed and far from the distribution of real speech in an ideal speech feature space, which makes discriminating straightforward and quick to learn categorization boundaries. Moreover, the scattered speech features also facilitate the subsequent identification of a specific spoofing method. Therefore, the triplet mining [22] method is ideal for the aforementioned causes since it can aggregate the bonafide speech feature while dispersing the spoofing ones.

$$L_{tri} = \sum_{i}^{N} \left(||f\left(x_i^a\right) - f\left(x_i^p\right)||_2^2 - ||f\left(x_i^a\right) - f\left(x_i^n\right)||_2^2 + \alpha \right), \tag{2}$$

where x_i^a, x_i^p, x_i^n represents the anchor sample, positive sample, and negative sample, respectively. The first term of Equation 2 computes the euclidean distance between the anchor and positive sample and the second term calculated the distance between the anchor and the negative one. By minimizing L_{tri}, the distance between the anchor and the positive sample will get closest while the anchor will be further away from the negative sample. we set α to 0.1 which is a margin value.

2.3. Total loss

In conclusion, the overall loss function L_{all} is described in Equation 3.

$$L_{all} = L_{cls} + \lambda_1 L_{ada} + \lambda_2 L_{tri}, \tag{3}$$

where L_{cls} utilize Binary Cross Entropy (BCE) to classify the feature in real or fake two categories. It's worth noting that weight normalization is used in fully conected layer (FC) of classifier which is widely utilized in face recognition [23, 24] to maintain the stability of the gradient descent and enhance the judgment of binary classification. λ_1 and λ_2 are set to 0.1 to balance the proportion of different loss functions.

3. Experiment

3.1. Database

To effectively employ DIRL strategy, the training data must encompass a wealth of attack types and be sourced from various domains. Thus, we select ASVspoof2019LA [25], WaveFake [26], and FakeAVCeleb [27] as our traing datasets. Specifically, ASVspoof2019LA is generated by 11 TTS and 8VC spoofing algorithms from VCTK [28]. WaveFake is an audio DeepFake dataset created by 7 spoofing methods from LJspeech [29]. FakeAVceleb is a multi-modal DeepFake dataset, the audio part is generated from Voxceleb [30]. All of the training data includes 26065 real utterances and 212035 fake utterances.

To test the generalization ability of our model, we try our best to collect the publicly available ADD datasets for evaluation. IWA [6], ASVspoof2021DF [31], JSUT [32], FAD [33] are used to evaluate our model. For FAD, we random select eight attacking type FC1-FC8, giving the following conditions: FC1-Lpcnet, FC2-stylegan, FC3-tacohifi, FC4-fasthifi, FC5-wavenet, FC6-straight, FC7-hifigan, FC8-pwg. FN1 and FN2 is the noise condition of FC1 and FC2. The test set condition is same as [33] which concludes 3500 spoofing and 3500 genuine utterances from the original domain.

3.2. Implementation details

All training audio files are resampled to 16kHz and trimmed or padded to 4s. For baseline Rawnet2 and AASIST, the input is the raw waveform of about 4s (64000 samples). For baseline XceptionNet, MesoInception, and Resnet18, we use 80-dimensional LFCCs with a shape of (80,404) as front-end. In all experiments, no DA techniques are used. During training, the parameters of W2V2 front-end are frozen. After front-end, we can get the last hidden states vector with shape of (201, 1024) as input of back-end. We use a fold setting similar as [34] but use 5 fold to cover all attack types. In each fold, we trained 5 epochs with validation. We divide the training set into 80% and the validation set into 20% and keep the attack type in validation of each fold not repeat.

4. Results and discussions

4.1. Feature experiments

To test the effect of single domain and cross-domain for different front-ends, we first train a single LCNN backbone without DIRL or DA strategy on ASVspoof2019LA train subset and test on ASVspoof2019LA evaluate subset, IWA, ASVspoof2021DF, FC1, and FN1. The results are shown in Table 1. W2V2 outperforms manual features for unknown domains in situations with small-scale data and limited attacks. Especially for FC1, despite FC1 is an invisible attack type for the training set and a cross-language evaluation, W2V2 can still achieve an EER of less than 5%. Manual features may perform well on 19eval dataset, but they encounter a sharp decrease in the out of domain. Thus, we utilize the W2V2 as our front-end.

Table 1: *EER(%) results for Feature expreiments. All systems are trained with ASVspoof2019 LA train subset.*

Feature	19LAeval	IWA	21DF	FC1	FN1
LFCC	1.82	43.12	27.83	32.91	34.40
W2V2	**0.63**	**24.50**	**8.07**	**4.48**	**26.85**

4.2. Out of domain experiments

To improve the generalization ability of ADD model, we use ASVspoof2019LA, WaveFake, and FakeAVCeleb three different datasets as our training data. The EER of the baseline and our model has shown in Table 2. W2V-ASDG represents our proposed method and achieves the best performance in 4.60% average EER. For IWA dataset, the original paper proposed that the current ADD model lacks generalization, and their model has a test EER for more than 40%. In our practice, even if training set is expanded to three domains, the model which utilizes handcrafted features and raw audio as front-end does not perform well on IWA. W2V2 front-end effectively reduces the EER to a single-digit level. For ASVspoof2021 DF dataset, owing to different compression, bitrate, and origin domain. The top of team only achieves 15.63% EER for DF subset. To the best of our knowledge, our proposed model with 2.22% EER is the lowest EER reported for the ASVspoof2021 DF database. For FAD and JSUT databases, our proposed ASDG method with W2V2 front-end gets the lowest EER in each condition. It is noteworthy that our training data only included English datasets, while the FAD and JSUT databases consist of Chinese and Japanese datasets respectively, creating a cross-language

Table 2: *EER(%) results for out of domain.*

Model	Features	IWA	21DF	JSUT	FC1	FC2	FC3	FC4	FC5	FC6	FC7	FC8	FN1	FN2	AVG
XceptionNet [35]	LFCC	60.29	31.94	26.43	8.34	18.97	2.71	11.48	30.45	37.74	12.57	2.08	49.48	35.02	25.19
MesoInception [36]	LFCC	66.10	34.85	21.20	14.17	8.68	2.54	19.34	24.60	46.71	27.20	2.17	44.65	34.77	26.69
Resnet18 [37]	LFCC	62.14	29.78	24.10	10.65	9.48	6.02	7.05	19.65	30.14	9.94	3.68	47.97	36.62	22.86
Rawnet2 [4]	Audio	32.74	21.53	47.60	28.25	29.57	33.77	63.31	38.31	49.14	31.48	36.33	33.06	32.48	36.74
AASIST [5]	Audio	19.38	12.94	6.95	4.25	7.22	14.14	10.80	7.25	24.60	18.94	9.05	28.65	29.85	14.92
ASDG	LFCC	52.95	37.80	12.14	1.82	6.25	1.92	6.77	7.56	40.98	5.62	1.55	40.92	35.89	19.40
ASDG (w/o) tri&ada	W2V2	7.70	3.99	6.25	3.01	1.91	3.02	10.82	1.40	5.20	11.05	1.88	27.96	18.85	7.93
ASDG (w/o) tri	W2V2	6.45	8.68	5.20	1.83	1.35	1.86	4.31	1.15	3.12	3.71	1.34	19.97	15.22	5.71
ASDG (w/o) ada	W2V2	6.66	3.28	5.14	1.68	1.82	1.94	4.14	1.34	2.89	3.68	1.22	19.23	16.17	5.32
W2V-ASDG	W2V2	**5.16**	**2.22**	**4.32**	**1.48**	**1.25**	**1.74**	**3.79**	**1.01**	**2.11**	**3.62**	**1.14**	**17.54**	**14.36**	**4.60**

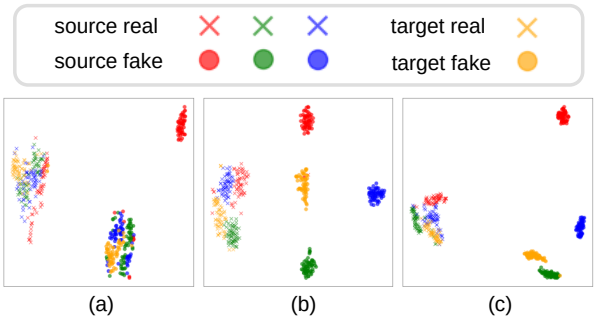

Figure 2: *Visualization of the feature space. The graphs of (a), (b), and (c) exhibit the T-SNE visualization of ASDG (w/o) tir&ada, ASDG (w/o) ada, ASDG respectively. Different colors indicate features from different domains: red=ASVSpoof2019LA, green=WaveFake, blue=FakeAVCeleb, yellow=Target. Different shapes represent different categories information: cross=real, point=fake.*

evaluation scenario. Among FC1-FC8, FC1-FC4 are spoofing methods that were not seen in the training database, yet the ASDG method still obtained consistency improvements. For N1-N2, the traditional feature extraction method (LFCC) was not able to clearly distinguish between real and fake speech under noise disturbance, while the W2V2 features showed obvious anti-noise ability even without DA.

4.3. Ablation Study

To verify the importance of the triplet mining and the single-side adversarial learning, the ablation study is added in the bottom of Table 1. ASDG (w/o) tir&ada represents the ADD system only using a feature generator using a classifier and BCE to identify the speech. ASDG (w/o) tri and ASDG (w/o) ada adopt adding single-side adversarial alone and adding triplet mining alone, respectively. Compared to the ASDG (w/o) tir&ada, the adversarial or triplet mining method can reduce the average EER by about 2%. W2V-ASDG is our proposed model which has the lowest EER of 4.60%.

4.4. Visualization Of The Feature Space

To analyze the feature space learned by our model and verify the effectiveness of the triplet mining and adversarial learning, we visualize the distribution of different features using T-SNE [38]. As shown in Figure 2, we randomly select 420 samples from four data domains, representing source domains 1-3 and the target domain respectively. In each domain, we select 60 samples for real utterances and 60 for fake utterances. In Figure 2(b), after adding the triplet loss, the distribution of fake utterances begins to separate into three different domains and far from the real utterances. In Figure 2(c), with the help of triplet mining and domain adversarial classifier, the distribution from different domains of genuine speech become interleaved and mixed together. Moreover, for unseen domain FC1, the distribution of fake speech forms a new cluster effectively differentiated from other source domains which may allow us to conduct a study on deepfake attribution.

5. Conclusions

In this paper, we introduce a generalized ADD system that incorporates a self-supervised front-end and a backbone utilizing the ASDG strategy. Specifically, we use W2V2-XLS-R as the front-end to extract the hidden states of the raw audio. Then, a modified version of LCNN as the backbone of the network is used to distinguish real or fake. In addition, Triplet mining method is designed to separate the fake speech and aggregate the real speech. Meanwhile, the single-side domain discriminator makes the real speech from different domains undistinguishable for the further aggregation of real speech. In this way, we learn a self-supervised domain-invariant feature representation to improve the generalization ability.

Our ADD model differs from current models as it prioritizes classification performance on the target domain rather than intra-domain performance. We expect to generate an ideal ADD feature space in a limited dataset and achieve that goal by visualization of the feature space. The test results reveal that our proposed model has the best performance across all conditions. Future work will concentrate on more application of DIRL method and transfer learning for ADD task.

6. Acknowledgements

The authors would like to thank the anonymous reviewers for their insightful comments. This work was supported by the Natural Science Foundation of China under grant No. 62201524, No. 62271455, No. 61971383 and the Fundamental Research Funds for the Central Universities under grant No. CUC22GZ002, No. CUC18LG024.

7. References

[1] A. Khan, K. M. Malik, J. Ryan, and M. Saravanan, "Voice spoofing countermeasures: Taxonomy, state-of-the-art, experimental analysis of generalizability, open challenges, and the way forward," *arXiv preprint arXiv:2210.00417*, 2022.

[2] A. Cohen, I. Rimon, E. Aflalo, and H. Permuter, "A study on data augmentation in voice anti-spoofing," *Speech Communication*, vol. 141, pp. 56–67, 2022.

[3] H. Tak, M. Kamble, J. Patino, M. Todisco, and N. Evans, "Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6382–6386.

[4] 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.

[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] N. Müller, P. Czempin, F. Diekmann, A. Froghyar, and K. Böttinger, "Does Audio Deepfake Detection Generalize?" in *Proc. Interspeech 2022*, 2022, pp. 2783–2787.

[7] Y. Zhang, G. Zhu, F. Jiang, and Z. Duan, "An Empirical Study on Channel Effects for Synthetic Voice Spoofing Countermeasure Systems," in *Proc. Interspeech 2021*, 2021, pp. 4309–4313.

[8] 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.

[9] G. Lavrentyeva, A. Novoselov, S., M. Volkova, A. Gorlanov, and A. Kozlov, "Stc antispoofing systems for the asvspoof2019 challenge," *arXiv preprint arXiv:1904.05576*, 2019.

[10] X. Wang and J. Yamagishi, "Investigating self-supervised front ends for speech spoofing countermeasures," *arXiv preprint arXiv:2111.07725*, 2021.

[11] Y. Xie, Z. Zhang, and Y. Yang, "Siamese network with wav2vec feature for spoofing speech detection." in *Interspeech*, 2021, pp. 4269–4273.

[12] Y. Eom, Y. Lee, J. S. Um, and H. R. Kim, "Anti-Spoofing Using Transfer Learning with Variational Information Bottleneck," in *Proc. Interspeech 2022*, 2022, pp. 3568–3572.

[13] J. W. Lee, E. Kim, J. Koo, and K. Lee, "Representation Selective Self-distillation and wav2vec 2.0 Feature Exploration for Spoof-aware Speaker Verification," in *Proc. Interspeech 2022*, 2022, pp. 2898–2902.

[14] J. M. Martín-Doñas and A. Álvarez, "The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9241–9245.

[15] 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," *arXiv preprint arXiv:2202.12233*, 2022.

[16] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[17] T. Matsuura and T. Harada, "Domain generalization using a mixture of multiple latent domains," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 749–11 756.

[18] L. Zhou, J. Luo, X. Gao, W. Li, B. Lei, and J. Leng, "Selective domain-invariant feature alignment network for face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 5352–5365, 2021.

[19] Z. Wang, Z. Wang, Z. Yu, W. Deng, J. Li, T. Gao, and Z. Wang, "Domain generalization via shuffled style assembly for face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4123–4133.

[20] W. Chen, X. Xing, X. Xu, J. Pang, and L. Du, "Speechformer: A hierarchical efficient framework incorporating the characteristics of speech," *arXiv preprint arXiv:2203.03812*, 2022.

[21] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[22] F. Schroff, D. Kalenichenko, and Philbin., "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[23] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[24] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[25] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019.

[26] J. Frank and L. Schönherr, "Wavefake: a data set to facilitate audio deepfake detection," *arXiv preprint arXiv:2111.02813*, 2021.

[27] H. Khalid, S. Tariq, M. Kim, and S. Woo, "Fakeavceleb: a novel audio-video multimodal deepfake dataset," *arXiv preprint arXiv:2108.05080*, 2021.

[28] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

[29] K. Ito and L. Johnson, "The lj speech dataset," *https://keithito.com/LJ-Speech-Dataset/*, 2017.

[30] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[31] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch *et al.*, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *arXiv preprint arXiv:2210.02437*, 2022.

[32] R. Sonobe, S. Takamichi, and H. Saruwatari, "Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis," *arXiv preprint arXiv:1711.00354*, 2017.

[33] H. Ma, J. Yi, C. Wang, X. Yan, J. Tao, T. Wang, S. Wang, L. Xu, and R. Fu, "Fad: A chinese dataset for fake audio detection," *arXiv preprint arXiv:2207.12308*, 2022.

[34] P. Kawa, M. Plata, and P. Syga, "Attack Agnostic Dataset: Towards Generalization and Stabilization of Audio DeepFake Detection," in *Proc. Interspeech 2022*, 2022, pp. 4023–4027.

[35] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[36] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2018, pp. 1–7.

[37] X. Chen, Y. Zhang, G. Zhu, and Z. Duan, "Ur channel-robust synthetic speech detection system for asvspoof 2021," *arXiv preprint arXiv:2107.12018*, 2021.

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