INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Adversarial Diffusion Probability Model For Cross-domain Speaker Verification Integrating Contrastive Loss

Xinmei Su[1], *Xiang Xie*[*1], *Fengrun Zhang*[1], *Chenguang Hu*[1]

[1]School of Information and Electronics, Beijing Institute of Technology, Beijing, China

suxinmei2022@126.com, xiexiang@bit.edu.cn, 1259619642@qq.com, 3220200551@bit.edu.cn

Abstract

In speaker verification, performance degradation caused by domain mismatch has been a common problem as the test domain lies outside the training distribution. In this paper, we present a novel domain transfer network called Adversarial Diffusion Probabilistic Model (ADPM), to better alleviate this problem. More specifically, ADPM is used to transfer melspectrogram from the source domain into the target domain. To generate the melspectrogram, we propose to regard the diffusion model as the generator and a discriminator is employed for adversarial training. We also explore the contrastive learning objective to retain the context information of source domain. The generated and the original feature maps from the source domain are fed into the ResNet34 network jointly to construct cross-domain speaker verification. We evaluate the proposed techniques on VOiCES dataset, and our best model achieves a relative 8.94% Equal Error Rate (EER) drop compared to the previous adaption methods.

Index Terms: speaker verification, cross-domain, diffusion models, contrastive learning

1. Introduction

Automatic Speaker Verification (ASV) which can verify whether two utterances are spoken from a same speaker, has been widely used in daily life [1]. Extracting speaker embeddings by deep-learning networks has become the most commonly used method in ASV systems. However, the embeddings extracted by the speaker extractor depend on large scale of datasets for training and show poor performance on cross-domain situations when a domain shift happens between the source domain for training and target domain for testing. Large datasets are difficult and expensive to be annotated, and the speaker embedding network can not fix domain mismatch problems. Therefore, domain adaptation (DA) is proposed by researchers to solve the problem of domain mismatch between resource-rich labeled source domain and resource-poor unlabeled target domain [2].

Probabilistic Linear Discriminant Analysis (PLDA) adaptation for target domain are usually adopted in traditional methods for ASV domain adaptation tasks [3]. In recent years, deep-learning approaches, such as CycleGAN [4, 5], are proposed for DA problems. These models utilize a generative network to fix the input features, which can better adapt to the output features. However, CycleGAN requires two Generative Adversarial Networks (GAN) [6] which can map from target domain to source domain both forward and inversely. Training GANs requires vast cost of time and computational resources. For the above reason, researchers are exploring several methods that contains both adversarial loss and contrastive loss in GANs to obtain a one-way training for DA tasks, such as Contrastive Learning for Non-parallel Voice Conversion (CVC) [7] models in Voice Conversion (VC). In CVC, it only requires a one-way GAN structure, which reduces the difficulty for training GANs significantly. By using noise contrastive estimation (NCE) [8], CVC can preserve content information since it builds correspondence between source and target spectrograms. In this way, the computational cost is greatly inclined and the performance of generating spectrograms is also improved.

Furthermore, another family of generative models called Diffusion Probabilistic Models (DPM) [9] attracts the attention of academia. Diffusion models show great capability to model the distribution of input images [10] and the images quality generated by diffusion model are significantly better than that of GAN model. In the application of speech, DPM family shows significant results in the reconstruction of raw waveform [11, 12] and spectrograms [13].

Inspired by recent work, we propose an adversarial DPM integrating the contrastive learning against the domain shift speaker verification task called ADPM. First, a diffusion model is used to reconstruct the spectrograms of the target domain by using the source domain spectrograms as input. What's more, drawing on the idea of adversarial learning, a discriminator is introduced which can continuously game with the DPM generator to generate feature maps closer to the target domain. Finally, to better fit the distribution of target domain data, the NCE-based patch loss is adopted to achieve better performances for one-way reconstructions of target spectrograms. The generated spectrograms with the target domain distributions are trained together with the original source domain data by a ResNet34 network [14, 15] for speaker verification.

The paper is organized as follows. In Section 1, the background for cross-domain ASV, and the related work of the generation models and contrastive learning is briefly reviewed. In Section 2, the method proposed is described in detail. In Section 3, the datasets and experimental settings are stated. In Section 4, the experimental results and ablation studies are presented to prove the effectiveness of our method. Conclusions and future works are discussed in Section 5.

2. Method

The overall framework of our method proposed is described in Figure 1. The acoustic features and the generated features from the ADPM is fed into the speaker verification network to achieve the speaker-level classification. There are two ways of

This work was supported by National Nature Science Foundation of China (Grant No.62071039) and Beijing Nature Science Foundation (Grant No.L223033). * is the corresponding author.

10.21437/Interspeech.2023-1205

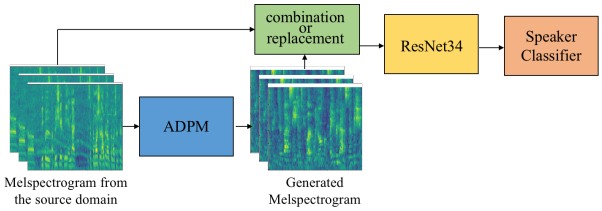

Figure 1: *The overall framework for cross-domain speaker verification.*

data usage sent to the neural network, one is the combination of generated features and source features, and the other is to replace the original source domain features with the generated features. The ADPM structure is presented in Figure 2. First, our DPM is an encoder-decoder network to generate features out-of domain. Second, the DPM is trained together with a discriminator to adopt adversarial learning for better fitting the distribution of the target domain. Third, the contrastive learning which reserves the content from the source utterances also help the network gain more effectiveness.

2.1. Diffusion probabilistic model

DPM posses a noise-adding forward process and a denoising reverse process that owns a powerful feature generation capability. Inspired by DPM in the filed of image generation, the DPM is put to use in our task for generating more feature maps as the front-end data-augmentation for cross-domain ASV.

The process of the DPM is followed by a Stochastic Differential Equation (SDE). In score-based [16] diffusion models, the SDE is writen as:

$$dX_t = \frac{1}{2}\Sigma^{-1}\left(\mu - X_t\right)\beta_t dt + \sqrt{\beta_t}dW_t, \quad t \in [0, T] \quad (1)$$

where the β_t follows a certain noise schedule. μ and Σ are the mean and variance of X_t. The forward process X_t can then be defined as:

$$\begin{aligned} X_t = &\left(I - e^{-\frac{1}{2}\Sigma^{-1}\int_0^t \beta_s ds}\right)\mu + e^{-\frac{1}{2}\Sigma^{-1}\int_0^t \beta_s ds}X_0 \\ &+ \int_0^t \sqrt{\beta_s}e^{-\frac{1}{2}\Sigma^{-1}\int_s^t \beta_u du}dW_s. \end{aligned} \quad (2)$$

By several forward diffusion processes, X_0 can be transformed into random variable X_t with the distribution above. That is, the SDE equation add noise into the images to achieve Gaussian noise eventually.

The reverse diffusion process is followed by an ordinary differential equation:

$$dX_t = \frac{1}{2}\left(\Sigma^{-1}\left(\mu - X_t\right) - \nabla \log p_t\left(X_t\right)\right)\beta_t dl \quad (3)$$

where the $\nabla \log p_t(X_t)$ denoted the log-density of the noisy data. Thus, the aim of the neural network of DPM is to train a network that can estimate $\nabla \log p_t(X_t)$. As noisy X_t can be sampled from X_0 by a Gaussian metric, the sampling formula can be defined as:

$$X_t = \rho(X_0, \sigma, \mu, t) + \varepsilon_t \quad (4)$$

The log-density of X_t given X_0 is calculated as:

$$\nabla \log p_{0t}(X_t|X_0) = -\lambda(\sigma, t)^{-1}\varepsilon_t \quad (5)$$

where $p_{0t}(X_t|X_0)$ follows the distribution of Guassian. For cross-domain speaker verification task, the input X_0 are clean filterbank features from the target domain. The whole structure of the DPM owns two components, an encoder $G_{DPM-enc}$ and a decoder $G_{DPM-dec}$. The encoder is an attention-based CNN structure [17] and the decoder is the traditional U-net [18]. The mean μ and variance Σ is computed by the encoder of DPM by feeding into the source domain data. The loss of the network can be computed as:

$$\mathcal{L}_{DPMt}(X_0) = \mathbb{E}_{\epsilon_t}[||s_\theta\left(X_t, t\right) + \lambda(\Sigma, t)^{-1}\epsilon_t||_2^2] \quad (6)$$

where ε_t is sampled from $\mathcal{N}\left(0, \lambda(\Sigma, t)\right)$.

2.2. Adversarial learning with DPM for melspectrogram generation

In tasks of domain adaptation, adversarial learning has a wide range of applications. For example, the basic theory of GAN [6] is to train a discrimiator and a generator. In generation process, it receives a random noise and generates features from the noise. In discrimination process, it determines if a picture is real. The result of the final game is: in the most ideal state, the generator can generate features that is enough to disguise the real feature. Learn from the structure of GAN, in this Section, the score-based DPM is utilized as the generator to generate the melspectrograms of the target domain with the source domain melspectrograms as input.

Furthermore, we additionally add a discriminator to distinguish the real image from the target domain and the fake image generated from the source domain. In this way, the diffusion model can gain better generation ability. The structure of our adversarial DPM model, which is named ADPM by us, is similar to GAN which has both a generator and a discriminator. Besides the normal objective in DPM, the generative and discriminative losses are also employed to help the model better differentiate the real melspectrograms in target domain from the melspectrograms generated from the source domain [19].Through the adversarial process of the generator and the discriminator, the DPM generator can generate more accurate target-domain melspectrograms.

The input spectral features follow a distribution $X = \{x \in \mathcal{X}\}$, where $\mathcal{X}$ is the distribution in the source domain. Similarly, the feature maps in the target domain follows another distribution, $Y = \{y \in \mathcal{Y}\}$. Following the idea, an adversarial loss is applied which can help the features of the input approach the target domain $\mathcal{Y}$. The adversarial loss is defined as:

$$\begin{aligned} \mathcal{L}_{\text{ADPM}}\left(G_{DPMX\to Y}, D_Y\right) = &\mathbb{E}_{\boldsymbol{y}\sim Y}\log D(\boldsymbol{y}) + \\ &\mathbb{E}_{\boldsymbol{x}\sim X}\log(1 - D(G_{DPM}(\boldsymbol{x}))) \end{aligned} \quad (7)$$

where $G_{DPM}(\boldsymbol{x})$ is the input features passed through the diffusion model, and D is the discriminator.

2.3. Contrastive learning for domain adaptation

Cross-domain speaker verification can be regarded as a task to map functions from source data to the target data. Adversarial learning can fix the problem of domain mismatch, such as CycleGAN [4] which innovatively put forward cycle-consistency objective to retain the content of the input image so that it can return to the input image itself during the secondary conversion. However, the function must be calculated forward and reversely

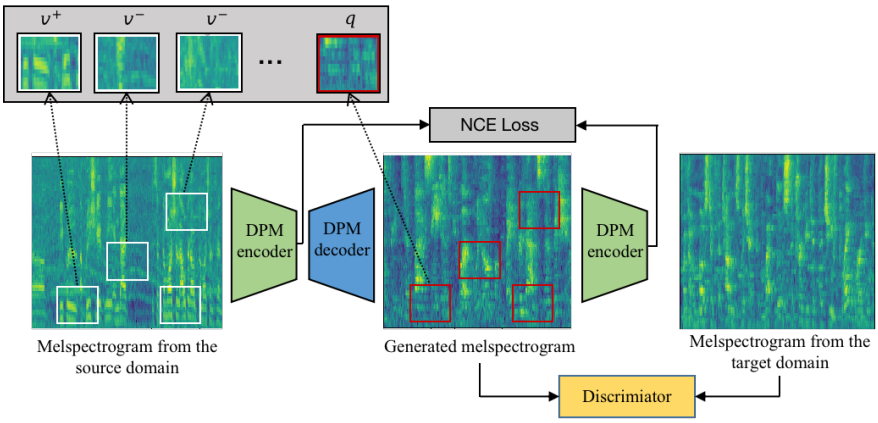

Figure 2: *The structure of our proposed ADPM.*

under a pixel-wise situation which consumes large computational resources.

In non-parallel generation tasks, well-generated features possess the content and speaker information of the source domain, and the distribution is similar to the target domain. The proposed ADPM model shows strong abilities to fit the distribution of the target domain, but has poor abilities to retain the content of the source domain. Therefore, the NCE [8] is adopted to preserve more information in the source domain. The main theory of NCE is to minimize the distance between the current small patch of image and its positive sample, and maximize the distance between the patch and its N negative samples. It is noted that the NCE function is only used in the encoder of DPM ($G_{DPM-enc}$). The objective of NCE is defined as:

$$\ell\left(\boldsymbol{q}, \boldsymbol{v}^{+}, \boldsymbol{v}^{-}\right)=-\log\left(\frac{\exp\left(\frac{\boldsymbol{q}\cdot\boldsymbol{v}^{+}}{\tau}\right)}{\exp\left(\frac{\boldsymbol{q}\cdot\boldsymbol{v}^{+}}{\tau}\right)+\Sigma_{n=1}^{N}\exp\left(\frac{\boldsymbol{q}\cdot\boldsymbol{v}_{n}^{-}}{\tau}\right)}\right) \tag{8}$$

where q means the query vector, v^+ and v^- are positive and negative samples respectively as illustrated in Figure 2. τ is the temperature parameter.

Inspired by [7], the query vectors are captured by small patches which can represent local information of the whole feature image. That is, the patches can represent the local information, such as phonemes, for better preserving the content information of the source domain. The method is set up by the assumption that the non-parallel correspondences is accumulated by certain local information. Thus, the NCE loss is computed based on small patches of last layer of encoder outputs. The final loss function is shown in Equation 9.

$$\mathcal{L}_{\mathrm{NCE}}\left(G_{DPM-enc}P, X\right)=\mathbb{E}_{\boldsymbol{x}\sim X}\sum_{n=1}^{N+1}\ell\left(\boldsymbol{q}^{n}, \boldsymbol{v}^{n}, \boldsymbol{v}^{(N+1)\backslash n}\right) \tag{9}$$

where n is the sample index of N negative samples.

Finally, the overall loss function of our method proposed is a weighted-sum of diffusion losses, adversirial loss and the NCE loss:

$$\mathcal{L}=\mathcal{L}_{\mathrm{DPM}}+\mathcal{L}_{\mathrm{ADPM}}+\mathcal{L}_{\mathrm{NCE}} \tag{10}$$

where the weight of each loss is set equally.

2.4. Speaker verification backend

After generating the spectral features with both the distribution of the target domain and the content information of the source domain, two different methods are proposed to make use of the generated images for SV tasks. The first approach let the mel-spectrograms feed into a conventional network for speaker verification together with the former features of the source domain. The second one is to substitute the original features with the generated features to prove the flexibility of our produced images. A ResNet34 [14] is trained as the back-end network to verify multiple speakers.

3. Experiments

3.1. Datasets

The source domain data used is the VoxCeleb [20] Dataset which is consist of the VoxCeleb1 [21] and VoxCeleb2 [22]. VoxCeleb1 contains 1211 speakers with 148642 utterances while VoxCeleb2 contains 5944 speakers with 1092009 utterances. All utterances from VoxCeleb are used for training. The Voices Obscured in Complex Environmental Settings (VOiCES) [23] Dataset is regarded as the target domain dataset which is totally different from the source domain. VOiCES comprises an Evaluation Set with 11392 utterances and a Development Set with 15904 utterances from 196 speakers. The Evaluation Set is the test set for verifying the performance of our methods. All utterances of the Development Set are served as the target domain resources, and the same number of utterances is randomly sampled from the VoxCeleb Dataset as the source domain resources, which aims to conduct the generation task proposed.

3.2. Experimental Settings

The 64-dimensional filterbank (Fbank) feature is extracted within a 25ms hamming window for every 10ms. No Voice Activity Detection (VAD) is adopted for Fbank features. No data augmentation is adopted in the whole experimental process. When training, the input features are cut into chunks that only contain a length of 4s (400 frames).

For adversarial learning task, the DPM encoder contains a 6-layer transformer encoder with two attention heads. Besides, the DPM decoder is the U-net [18] structure which is composed of a down-sampling and up-sampling ResNet net-

work. The discriminator is the Patch-Discriminator same as PatchGAN [24]. The iteration of the diffusion denoising process is set to 100. The learning rate of the Adam optimizer is 0.01. The cross-domain DPM-GAN is trained for $127k$ steps on single GPU (NVIDIA RTX 3090 with 24GB memory) with a batch size of 20.

For SV task, the training of the ResNet34 [14] is conducted on the ASV-Subtools [25] platform which is a PyTorch framework. The learning rate of the stochastic gradient decent (SGD) follows a ReduceLROnPlateau with a weight decay of $5e^{-4}$ and a initial learning rate of 0.02. The AMsoftmax [26] loss is employed with a scale of 30 and a margin of 0.2. For the back-end scoring, both LDA and PLDA is processed.

The Equal Error Rate (EER) and Minimum Detection Cost Function (minDCF) is the evaluation metric for our speaker verification task. The better the SV system is, the lower EER it gets. Moreover, the minDCF considers prior probability and different costs, which is more reasonable than EER. The p_{target} is set to 0.01 in minDCF for evaluation.

4. Results

Table 1 shows the results of our proposed ADPM structure compared with several robust systems recently proposed. DEAAN [27] and CDMA [28] are the best models for cross-domain speaker verification. As is seen in Table 1, ADPM achieves a relative EER improvement of 23.80% and 8.94% compared to DEAAN and CDMA respectively. This proves that the idea to generate more fbank features with the distribution of the target domain can enhance the back-end training for speaker verification task.

Table 1: *Comparison of methods before and our proposed method.*

Model	EER	minDCF
DEAAN [27]	5.21	0.394
CDMA [28]	4.36	0.369
ADPM (combination)	3.97	0.328

4.1. Ablation study

In this Subsection, we experiment certain ablation studies to prove the effectiveness of all the components proposed by us as shown in Table 2. The Baseline system is a simple ResNet34 speaker embedding extractor trained on the source domain. The eventual scheme of ADPM (combination) is to adopt joint training with both the source domain data and the generated data. One may argue that it increases the amount of data to make the supervised training of SV better. To address this problem, the source domain data is replaced by the generated Fbank features (ADPM (replacement)) to ensure the total utterances are the same as the Baseline. It can be seen from the result that even if the amount of data is not increased, our system shows a relative drop of 3.44% in EER compared with CDMA. Furthermore, the contrastive learning objective (NCE) is removed with a relative increase of 5.54% compared to the whole architecture proposed, which means that the contrastive learning can help the network preserve the content information of the source domain and the distribution of the target domain. Eventually, the adversarial learning objective is also eliminated (DPM). The EER raise to 4.47% which proves that using DPM as a generator together with a discriminator can boost the performance to better fit the source domain data to the target domain.

Table 2: *Ablation study of all components proposed in ADPM.*

Model	EER	minDCF
ADPM (combination)	3.97	0.328
ADPM (replacement)	4.21	0.346
ADPM - NCE (combination)	4.19	0.347
DPM (combination)	4.47	0.379
Baseline	6.62	0.548

4.2. Visualization of the generated features

As is seen in Figure 3, the most intuitive differences are the background noise and silent parts (external attributes). There is several continuous background noise in the source domain feature map (a), in contrast, the target domain Fbank feature (b) is relative cleaner and has more silent parts. The feature (c) and (d) generated from (a) is cleaner than the source filed. It can be shown that the distribution of the generated one is closer to the target domain data, while the content information of the source domain, such as pitch period and high pitch (speaker information), etc, is also preserved. The visualization of the acoustic features can prove the generation performance of our model.

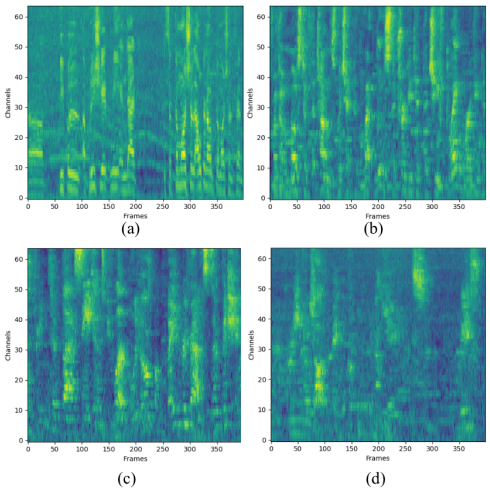

Figure 3: *A visualization for the acoustic features, where (a) is from the source domain, (b) is the target domain feature map, (c) is the generated feature with the input of the same utterance (a) by ADPM, and (d) is the generated feature from a random selected sample in source domain.*

5. Conclusions

In this paper, an adversarial DPM method with contrastive learning (ADPM) which can fit the source domain data to the target domain data to achieve better performance in domain shift speaker verification is proposed. First, a diffusion model is applied to generate acoustic features of the target domain. Second, adversarial learning objective is utilized to gain better performance in generation. Third, the NCE contrastive learning loss helps the network resume the content information of the source domain to avoid feature distortion. The model proposed achieves 8.94% decrease of EER. In future, we will make use of the joint learning to train the generation task and the speaker verification task simultaneously to acquire better performance in cross-domain speaker verification.

6. References

[1] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, pp. 1–22, 2004.

[2] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," *Advances in neural information processing systems*, vol. 19, 2006.

[3] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "Plda for speaker verification with utterances of arbitrary duration," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7649–7653.

[4] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[5] P. S. Nidadavolu, S. Kataria, J. Villalba, and N. Dehak, "Low-resource domain adaptation for speaker recognition using cycle-gans," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 710–717.

[6] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[7] T. Li, Y. Liu, C. Hu, and H. Zhao, "Cvc: Contrastive learning for non-parallel voice conversion," *arXiv preprint arXiv:2011.00782*, 2020.

[8] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.

[9] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*. PMLR, 2015, pp. 2256–2265.

[10] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[11] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," *arXiv preprint arXiv:2009.09761*, 2020.

[12] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "Wavegrad: Estimating gradients for waveform generation," *arXiv preprint arXiv:2009.00713*, 2020.

[13] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8599–8608.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 2016, pp. 630–645.

[15] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances." in *Interspeech*, 2017, pp. 1487–1491.

[16] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

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

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[19] Z. Wang, H. Zheng, P. He, W. Chen, and M. Zhou, "Diffusion-gan: Training gans with diffusion," *arXiv preprint arXiv:2206.02262*, 2022.

[20] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2019.

[21] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[22] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[23] C. Richey, M. A. Barrios, Z. Armstrong, C. Bartels, H. Franco, M. Graciarena, A. Lawson, M. K. Nandwana, A. Stauffer, J. van Hout, P. Gamble, J. Hetherly, C. Stephenson, and K. Ni, "Voices obscured in complex environmental settings (voices) corpus," 2018.

[24] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[25] F. Tong, M. Zhao, J. Zhou, H. Lu, Z. Li, L. Li, and Q. Hong, "ASV-Subtools: Open source toolkit for automatic speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6184–6188.

[26] 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.

[27] M. Sang, W. Xia, and J. H. Hansen, "Deaan: Disentangled embedding and adversarial adaptation network for robust speaker representation learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6169–6173.

[28] J. Li, J. Han, and H. Song, "Cdma: Cross-domain distance metric adaptation for speaker verification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7197–7201.