DOMAIN ROBUST DEEP EMBEDDING LEARNING FOR SPEAKER RECOGNITION

Hang-Rui Hu[1], *Yan Song*[1], *Ying Liu*[1], *Li-Rong Dai*[1], *Ian McLoughlin*[1,2], *Lin Liu*[3]

[1]National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, China.
[2]ICT Cluster, Singapore Institute of Technology, Singapore.
[3]iFLYTEK Research, iFLYTEK CO. LTD., Hefei, China.

ABSTRACT

This paper presents a domain robust deep embedding learning method for speaker verification (SV) tasks. Most recent methods utilize deep neural networks (DNN) to learn compact and discriminative speaker embeddings from large-scale labeled datasets such as VoxCeleb and the NIST SRE corpus. Despite the success of exiting methods, performance may degrade significantly for new target datasets, mainly due to the distribution discrepancy between training and test domains. Moreover, how corpora are collected, and the languages they contain differ, leading to them spanning multiple, perhaps mismatched, latent domains. To address this, a multi-task end-to-end framework is proposed to learn speaker embeddings from both labeled source and unlabeled target datasets. Motivated by label smoothing, a smoothed knowledge distillation (SKD) based self-supervised learning method is designed to exploit latent structural information from the unlabeled target domain. Furthermore, a domain-aware batch normalization (DABN) module aims to reduce the cross-domain distribution discrepancy, while a domain-agnostic instance normalization (DAIN) module aims to learn features that are robust to within-domain variance. Evaluation on NIST SRE16 demonstrates significant performance gains.

Index Terms— Speaker Verification, Unsupervised Domain Adaptation, End-to-End, Label Smoothing, Knowledge Distillation

1. INTRODUCTION

Speaker verification (SV) systems automatically determine whether a speech utterance belongs to a given speaker. Performance relies heavily on the compactness and discrimination capability of speaker embeddings, one aim of recent deep neural network systems that learn from large-scale labeled dataset in a supervised way [1, 2, 3, 4, 5, 6, 7].

Generally, the underlying assumption of supervised learning is that the training and test data are independent and identically distributed, but this can be untrue in practice due to the domain shifts. For example, in the NIST SRE16 evaluation [8], the training dataset is collected from the Switchboard corpus and previous NIST SRE evaluations, which contain utterances from English speakers, while the target-domain dataset contain utterances from unknown Cantonese and Tagalog speakers. In such scenarios, pre-trained SV systems may suffer significant performance degradation. It is challenging to develop domain robust deep learning methods for SV.

Given the unlabeled target-domain dataset, unsupervised domain adaptation (UDA) aims to align the statistics of the source and target distributions with a linear transformation for the probabilistic linear discriminative analysis (PLDA) backend [9, 10, 11, 12]. Several recent end-to-end based methods learned a nonlinear transformation to align correlations of layer activations in DNN [13, 14, 15], while some methods used adversarial training strategies to mitigate the domain mismatch [16, 17, 18, 19, 20]. These methods generally minimize domain discrepancy but neglecting speaker label information, which may affect alignment and reduce generalization performance. In [21], unsupervised clustering based domain adaptation was proposed to estimate pseudo-labels of target domain data and then finetune the PLDA backend. More recently, the self-supervised learning based domain adaptation (SSDA) method leveraged potential label information from the target domain and adapted the discrimination ability from the source domain simultaneously [22]. However, noisy estimated label information in the target domain may hinder any performance improvement.

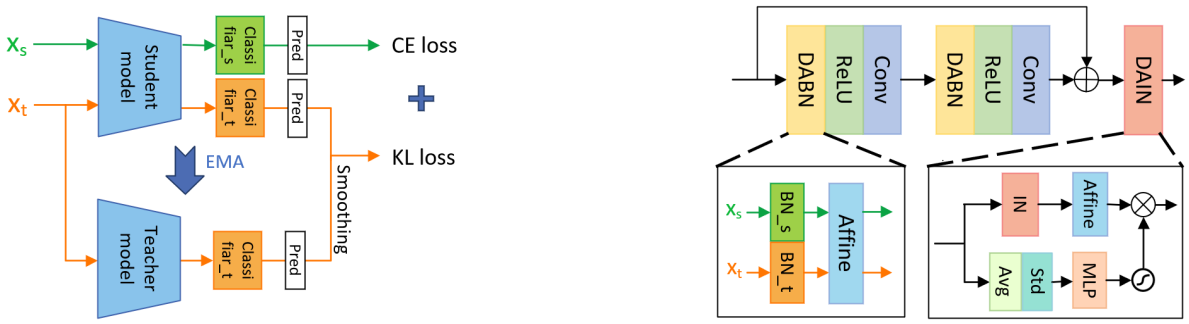
In this paper, we propose domain robust deep embedding learning for SV in which a multi-task framework, shown in Fig. 1a, learns speaker embeddings from both the labeled source domain and unlabeled target-domain datasets. Similar to [22], we assume each utterance contains only one speaker in short time duration. Instead of performing contrastive learning to construct positive and negative pairs, we propose a smoothed knowledge distillation (SKD) method, exploiting both label smoothing regularization (LSR) [23] and knowledge distillation (KD) [24] to effectively reduce label noise and exploit latent knowledge from the target domain. Specifically, SKD minimizes the Kullback-Leibler (KL) divergence between student and smoothed teacher outputs, as detailed later in Section. 3.1.

We furthermore design domain robust modules, domain-aware batch normalization (DABN) and domain-agnostic instance normalization (DAIN), to improve modeling capability on one domain as well as its generalization to other domains, without fine-tuning. Specifically, DABN performs targeted batch normalization (BN) for source and target domains, designed to reduce cross-domain distribution discrepancies. DAIN combines instance normalization (IN) with a channel attention mechanism to learn features that are robust to within-domain variance. DABN and DAIN are integrated in a Residual block, as shown in Fig. 1b. We evaluate the effectiveness of these methods on the NIST SRE16 benchmark, yielding a significant EER improvement of 19% relative to baseline.

2. OVERVIEW OF THE PROPOSED FRAMEWORK

In the proposed multi-task SKD based framework, shown in Fig. 1a, source labeled data is sent to the student model directly and supervised by cross entropy (CE) loss with one-hot ground truth labels. For unlabeled target domain data, we first assume each sample is an

Yan Song is the corresponding author. This work was supported by the Leading Plan of CAS (XDC08030200)



(a)Smoothed Knowledge Distillation based learning method

(b) Residual block with domain-robust modules

Fig. 1. The multi-task framework for both source and target domain datasets. (a) SKD based self-supervised learning. The student is jointly trained on the labeled source and unlabeled target-domain datasets via self-supervised learning. (b) Residual block with domain-robust modules. DABN uses different BNsto exploit source and target domains to reduce cross-domain discrepancy while DAIN uses mean and variance information to introduce a channel attention mechanism.

individual category and align each to a unique pseudo label. Note that the initial pseudo labels can also be estimated by some unsupervised learning methods. The target domain data is then fed into the teacher and student models simultaneously, and supervised by SKD loss-the Kullback-Leibler divergence (KL) between the student output and the smoothed teacher output. The total loss can be writen as the weighted sum of the losses from both domains:

$$\mathcal{L}_{total} = \mathcal{L}_{CE}^{(source)} + \lambda_t \cdot \mathcal{L}_{SKD}^{(target)} \tag{1}$$

Note that the source and target domains, with different numbers of categories, don't share the classifier. Also notice that the student model parameters are updated by stochastic gradient descent, whereas those of the teacher are obtained by accumulating the parameters of the student model every iteration through an exponential moving average (EMA) strategy.

Our experiments use ResNet-18 with attentive aggregation to extract features in both teacher and student models. Allowing source and target domains to share one student feature extractor works because both domains are well correlated. However, sharing domain-specific information such as BN[1] statistics, across two domains is inappropriate when the domain gap is large. We design a domain-aware batch normalization (DABN) module to replace all BNs in the network to separate such information. As shown in Fig. 1b, DABN consists of two branches, one for the source domain and another for the target domain. During training, data from different domains is whitened with their own statistics, and then passed through a shared affinity to transform to the same distribution. Moving averages of the global mean and variance for both domains are separately accumulated. During test, samples from each domain are whitened using the corresponding global statistics. In this way, domain-relevant information within the network can be effectively removed, allowing the network to better learn domain-invariant features.

IN is known as a kind of style normalization, and has been widely used to eliminate image style or domain information to improve model generalization [25, 26], However, IN completely discards the mean and variance statistics of samples, which degrades system discrimination capability [27, 28]. In this paper, we design a domain-agnostic IN (DAIN) module to alleviate the discrepancy among domains, while recovering useful information from the removed statistics. As shown in Fig. 1b, we construct an additional branch comprising two FC layers, taking the discarded mean and standard deviation of samples as input. Its output is then passed through a sigmoid activation to obtain attention weights, which we then multiply with the IN output to as an attention mechanism. DABN and DAIN together help the proposed SKD method learn robust domain-invariant features, and can be easily applied to many advanced domain adaptation methods.

[1]Our BN differs slightly from convention which considers the normalization and channel-wise affine transformation as separate parts.

3. METHOD

3.1. Smoothed knowledge distillation (SKD)

Inspired by label smoothing regularization (LSR) [23] and knowledge distillation (KD) [24], we proposed a novel loss function we name smoothed knowledge distillation (SKD) to reduce the label noise and exploit the latent structural information of target domain.

LSR and KD are two common learning strategies using soft labels which are complementary to each other [29, 30]. LSR reduces the confidence of the hard label, but divides it equally among all negative classes, which may not always be appropriate. KD learns inter-class relationships captured by the teacher model, but cannot solve overfitting of noisy labels.

Formally, given the target domain input with the pseudo label c, the output predictions for category k of the student and teacher model are denoted as $p(k)$ and $q(k)$, respectively, then the SKD loss can be written as:

$$\mathcal{L}_{SKD}(p, q, c) = \mathcal{L}_{KL}(p, q') \tag{2}$$

$$q'(k) = \begin{cases} \gamma \cdot q(k) + \beta, & k = c \\ A \cdot q(k)^{1/t}, & k \neq c \end{cases} \tag{3}$$

where q' is the smoothed teacher output, and $A = \frac{1-q'(c)}{\sum_{k \neq c} q(k)^{1/t}}$ is the normalization coefficient that ensures q' will sum to 1. γ and β are hyper-parameters that adjust the confidence of the main category, Note that $\gamma + \beta$ can be less than 1, in which case the overconfidence in a noisy label can be explicitly alleviated. As for negative classes, temperature t is used to smooth the output prediction of the teacher model only, to affect the degree of reliance on the inter-class relationship learned by teacher model.

If the teacher model is not credible, a smaller γ (for main category) and a larger t (for negative categories) can be chosen. Specifically, when $\gamma = 0$ and $t = \infty$, SKD becomes a standard LSR loss. Furthermore, note that when $\beta = 1$ in this case, it becomes a standard CE loss.

3.2. Domain-aware batch normalization (DABN)

DABN explicitly aligns the first and second moments of different domains in each middle layer without extra computation. It consists of domain-specific standard BN branches and a shared affine layer (see Fig. 1b). Data from different domains are whitened with their own statistics, and then transformed to the same distribution with domain-shared affine parameters γ and β.

Let $x^d \in \mathbb{R}^{N\times C\times F\times T}$ be an input mini-batch belonging to domain d on a certain layer, then DABN modules can be written as:

$$y = \hat{x}^{(d)} \cdot \gamma + \beta \tag{4}$$

$$\hat{x}^{(d)} = \left(x^{(d)} - \mu_{BN}^{(d)}\right) / \sqrt{\sigma_{BN}^{2(d)} + \epsilon} \tag{5}$$

During training, the domain-specific mean $\mu_{BN}^{(d)}$ and variance $\sigma_{BN}^{2(d)}$ along the channel dimension are computed in a mini-batch, while during inference, the global statistics exponential moving average is used to normalize features.

3.3. Domain-agnostic instance normalization (DAIN)

DAIN alleviates the discrepancy among latent sub-domains while recovering useful statistics otherwise discarded by IN, to maintain discrimination ability.

Let $x \in \mathbb{R}^{N\times C\times F\times T}$ be an input mini-batch to a certain layer, then DAIN modules can be written as:

$$y = (\hat{x} \cdot \gamma + \beta) \cdot \alpha \tag{6}$$

$$\hat{x} = \frac{x - \mu_{IN}}{\sqrt{\sigma_{IN}^2 + \epsilon}} \tag{7}$$

$$\alpha = f\left([\mu_{IN}, \sigma_{IN}]\right) \tag{8}$$

where $[,]$ is the concatenation operation, μ_{IN} and σ_{IN} are calculated in the same way as BN, but for each individual sample instead of mini-batch.

In this case, the attention branch can be expressed as

$$f(\cdot) = \sigma(W_2\delta(W_1(\cdot)) \tag{9}$$

where $W_1 \in \mathbb{R}^{2C\times\frac{C}{r}}, W_2 \in \mathbb{R}^{\frac{C}{r}\times C}$ are FC layer parameters, δ is the ReLU function, σ is the sigmoid function. r is the reduction rate, which is empirically set to 2 in experiments.

4. EXPERIMENTS

Experimental setup: Experiments are conducted on the NIST-SRE 2016 evaluation, including [8], which incorporates Tagalog and Cantonese telephone speech. The training data primarily consists of telephone speech from past issues of NIST-SRE (2004-2010) plus Switchboard. In addition, an unlabeled set of 2272 recordings in the development set is provided to adapt systems.

The feature extraction process uses the Kaldi toolkit [11]. In our implementation, 41-dimensional filter bank outputs (FBank) are used as acoustic features, obtained from 25ms windows with 10ms shift between frames. We apply mean-normalization over a sliding window of 3s, and use voice activity detection (VAD) to remove silent segments. The features from the training set are randomly truncated into short slices ranging from 2 to 4s.

Model configuration: The baseline model uses the ResNet-18 backbone as in [31], with the number of heads in the attentive bilinear pooling set to 8. The scaling factor after L2-norm is related to the number of classes, and this is set to 30 for the source domain (over 5000 classes) and 20 for the target domain (2272 classes), and the loss weight λ_t in Eq.(1) is set to 1.5 to stabilize the back-prop gradient. The momentum parameter in EMA is set to 0.95 for the parameters update of the teacher model.

All neural networks are implemented using the PyTorch framework. The mini-batch for training is set to 256, including 32 target domain data items and 224 source domain data items. The networks are optimized using stochastic gradient descent (SGD) with momentum of 0.9 and weight decay of 5e-4, and are trained for 60 epochs with initial learning rate of 0.1, gradually decreasing to 0.0001.

To demonstrate the effectiveness of the proposed method, we compared the performance of several systems as follows.

Baseline: ResNet-18 trained on the source domain training set, and evaluated on the target domain test set without UDA.

DABN: All BNs in the **Baseline** system are replaced with DABNs.

DABN+DAIN: This system builds on the **DABN** system by adding DAINs at the end of each residual block.

DABN+SKD: This system implements the proposed SKD method under a teacher-student framework. All BNs in the networks are replaced with DABNs, with the γ, β and t in SKD loss are set to 0.5, 0.5 and 10 respectively.

DABN+CE: The γ and β in the **DABN+SKD** system are set to 0 and 1, which makes the SKD loss degenerate into a standard CE.

DABN+LSR: The γ, β and t in the **DABN+SKD** system are set to 0, 0.6 and ∞ respectively, which makes the SKD loss degenerate into a standard LSR.

DABN+DAIN+SKD: This system is implemented on the **DABN+SKD** system by adding a DAIN at the end of each residual block. The γ, β and t in the SKD loss are set to 0.4, 0.4 and 100 respectively.

DABN+DAIN+LSR: The γ, β and t in the **DABN+DAIN+SKD** system are set to 0, 0.5 and ∞ respectively.

It is worth noting that, in the **+CE** and **+LSR** systems, the teacher model does not actually operate. Hence the smoothed target could be manually designed instead of being output by the teacher.

4.1. Experimental results

The main results are reported in Table 1. It is evident that the proposed system outperforms the baseline by a large margin. Specifically, the **DABN+CE** system achieves an EER of 14.59%, which is worse than the 13.85% of the baseline. This indicates that, under our assumption that each target domain sample belongs to an individual category, which is adopted by many unsupervised contrastive learning methods, the hard one-hot pseudo label introduces heavy noise, making the system overconfident. Furthermore, this assumption results in only one sample per category, in which case the CE loss is no longer suitable as an optimization objective.

Thanks to the label noise reduction capacity of LSR, the **DABN+LSR** system achieves an EER of 12.12%, compared to 13.85% by the baseline. But EER can reduce to 11.94% when employing the proposed SKD method. When inserting the DAIN modules into the system, EER further reduces to 11.22%. This result is a 19% relative improvement over the ResNet-18 baseline system, demonstrating the effectiveness of our proposed method for robust domain-invariant feature learning. It is worth noting that the

Table 1. Cosine EER(%) results of the comparison systems on NIST-SRE 2016 evaluation.

System	Pooled	Tagalog	Cantonese
Baseline	13.85	18.95	8.73
DABN+CE	14.59	19.56	9.73
DABN+LSR	12.12	16.63	7.57
DABN+SKD	11.94	16.45	7.43
DABN+DAIN+LSR	11.45	15.48	7.35
DABN+DAIN+SKD	11.22	15.38	7.16

deeper and wider ResNet-34 may achieve better baseline result, and consistent performance improvement can be obtained. Due to the computational complexity, we leave it to the future work.

Comparison of domain-robust modules – We study the effect of the proposed DABN and DAIN modules, in Table 2. In the Kaldi recipe, the extracted embeddings should subtract the global mean embedding before computing verification scores. We found it is significant to align the embeddings with the corresponding global mean for each domain, when there is a domain mismatch. To illustrate this point and explain the role of DABN, we also reported the results obtained by computing scores *without* the mean subtraction operation, termed as **Baseline w/o mean_sub** in Table 2. Note that none of the systems listed in Table 2 made use of the proposed SKD method.

Table 2. Cosine EER(%) results of DABN and DAIN on the NIST-SRE 2016 evaluation, without use of SKD.

System	Pooled	Tagalog	Cantonese
Baseline	13.85	18.95	8.73
Baseline w/o mean_sub	20.27	24.55	12.54
DABN	13.09	17.37	8.36
DABN+DAIN	12.25	16.81	7.86

First, we can see that EER is reduced from 20.27% to 13.85% after the mean alignment, indicating that there is a large mean discrepancy between the distributions of the source and target domains, so that sharing the mean and variance statistics is improper for those two domains. When we replaced all the BNs in the network with DABNs to separate the statistics of each domain, in the feature extraction phase, the samples from each domain are whitened using the corresponding global statistics. In this case, the system achieves an EER of 13.09% without the mean subtraction operation, which is slightly better than the baseline system. This illustrates that the DABN can help the feature extractor implement the mean and variance alignment.

When inserting the DAIN modules into the system, EER is further reduced to 12.25%. This indicates that DAIN can effectively alleviate the discrepancy among domains without hurting the discrimination capability of the system.

Comparison with existing systems – To further evaluate the effectiveness of the proposed method, we compare the cosine EER results to state-of-the-art systems using the same dataset in Table 3. **DANSE** adopted an adversarial training strategy with a gradient reverse layer (GRL). **FuseGan** was the ensemble of three GAN-based models, of which **LSGAN** was the best single model. We can see from the results that our method can achieve better cosine EER than either DANSE and LSGAN for the same conditions, although FuseGan achieves the overall best performance due to its model ensemble strategy. We believe that our system could outperform FuseGAN if it were similarly fused with other models.

Table 3. Cosine EER(%) comparison with other end-to-end adaptation methods on NIST-SRE 16 evaluation.

System	Pooled	Tagalog	Cantonese
DANSE [16]	13.29	17.87	8.84
LSGAN [17]	11.74	15.63	7.90
FuseGAN [17]	10.88	14.84	6.93
DABN+DAIN+SKD (ours)	11.22	15.38	7.16

Table 4. PLDA EER(%) comparison with other end-to-end adaptation methods on NIST-SRE 16 evaluation.

System	Similarity	Pooled	Tagalog	Cantonese
WGAN [18]	PLDA	13.25	19.12	7.39
	APLDA	9.42	-	-
MMD [15]	PLDA	9.03	-	-
	APLDA	8.29	-	-
Baseline (ours)	PLDA	10.79	15.76	5.49
	APLDA	7.50	11.26	3.60
SKD+DABN	PLDA	9.81	14.37	5.26
+DAIN (ours)	APLDA	6.95	10.13	3.64

PLDA results – In addition to the above evaluations, we compare the PLDA results of our system with other end-to-end adaptation systems in Table 4, where **WGAN** used Wasserstein GAN loss to alleviate the domain discrepancy, **MMD** minimized domain-wise MMD loss on multiple layers, and APLDA refers to Kaldi's adaptive PLDA [11].

When we further apply adaptation on PLDA, the EER result of our system can be reduced from 9.81% to 6.95%, showing that the proposed method, which implements domain adaptation at the front-end, can be easily incorporated with back-end adaptation.

Besides, notice that compared to baseline, the APLDA improvement of our proposed system (EER from 7.50% to 6.95%) is not as significant as the cosine improvement (EER from 13.85% to 11.22%). According to the cosine improvement, we believe that the discrimination and robustness of the extracted embeddings have been strengthened. The main reason for the above situation is that the amount of the target domain data is too small because we did not use any data augmentation strategy, making it difficult to greatly improve the inter-class and intra-class covariance of PLDA [12]. Data augmentation strategies or other back-end adaptation methods may be further combined with our work to achieve better performance.

5. CONCLUSION

This paper has presented a novel domain robust deep embedding learning method for SV – an end-to-end trained framework that learns speaker embeddings from both labeled and unlabeled target datasets in a smoothed knowledge distillation (SKD) framework. We also develop domain-aware batch normalization (DABN) to reduce cross-domain distribution discrepancies, and domain-agnostic instance normalization (DAIN) to learn features that are robust to within-domain variance. Meanwhile, we do not require data augmentation to generate new target domain data, or use adversarial or contrastive loss, and require very little computational overhead. The proposed methods are evaluated extensively on NIST SRE16, demonstrating significant performance gains over baseline. We believe these methods can in future be combined with complimentary approaches to further significantly improve results.

6. REFERENCES

[1] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, et al., "Deep neural networks for extracting Baum-Welch statistics for speaker recognition," in *The Speaker and Language Recognition Workshop*, 2014, pp. 293–298.

[2] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proc. Interspeech*, 2017, pp. 999–1003.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, et al., "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[4] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu, "Deep speaker: an end-to-end neural speaker embedding system," in *Interspeech*, 2017.

[5] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer, "End-to-end text-dependent speaker verification," in *Proc. ICASSP*, 2016, pp. 5115–5119.

[6] Weicheng Cai, Jinkun Chen, and Ming Li, "Analysis of length normalization in end-to-end speaker verification system," in *Proc. Interspeech*, 2018.

[7] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[8] Seyed Omid Sadjadi and etc Kheyrkhah, "The 2016 NIST speaker recognition evaluation.," in *Interspeech*, 2017, pp. 1353–1357.

[9] Baochen Sun, Jiashi Feng, and Kate Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016, vol. 30.

[10] Kong Aik Lee, Qiongqiong Wang, and Takafumi Koshinaka, "The coral+ algorithm for unsupervised domain adaptation of PLDA," in *ICASSP 2019*. IEEE, 2019, pp. 5821–5825.

[11] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, and N. Goel et al, "The kaldi speech recognition toolkit," in *ASRU*, 2011.

[12] Bengt J. Borgstrom, Elliot Singer, Douglas Reynolds, and Omid Sadjadi, "Improving the effectiveness of speaker verification domain adaptation with inadequate in-domain data," in *Interspeech*, 2017, pp. 1557–1561.

[13] Baochen Sun and Kate Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European conference on computer vision*. Springer, 2016, pp. 443–450.

[14] Wei-Wei Lin, Man-Wai Mak, et al., "Reducing domain mismatch by maximum mean discrepancy based autoencoders.," in *Odyssey*, 2018, pp. 162–167.

[15] Weiwei Lin, Man-Mai Mak, Na Li, Dan Su, and Dong Yu, "Multi-level deep neural network adaptation for speaker verification using MMD and consistency regularization," in *ICASSP 2020*. IEEE, 2020, pp. 6839–6843.

[16] Gautam Bhattacharya, Jahangir Alam, and Patrick Kenny, "Adapting end-to-end neural speaker verification to new languages and recording conditions with adversarial training," in *ICASSP 2019*. IEEE, 2019, pp. 6041–6045.

[17] Gautam Bhattacharya, Joao Monteiro, et al., "Generative adversarial speaker embedding networks for domain robust end-to-end speaker verification," in *ICASSP 2019*. IEEE, 2019, pp. 6226–6230.

[18] J. Rohdin, T. Stafylakis, A. Silnova, H. Zeinali, and O. Plchot, "Speaker verification using end-to-end adversarial language adaptation," in *ICASSP 2019*, 2019.

[19] Wei Xia, Jing Huang, and John HL Hansen, "Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation," in *ICASSP 2019*. IEEE, 2019, pp. 5816–5820.

[20] S. Kataria, J. Villalba, et al., "Deep feature cyclegans: Speaker identity preserving non-parallel microphone-telephone domain adaptation for speaker verification," in *INTERSPEECH*, 2021, pp. 1079–1083.

[21] Stephen H Shum, Douglas A Reynolds, Daniel GarciaRomero, and Alan McCree, "Unsupervised clustering approaches for domain adaptation in speaker recognition systems," in *Proc. of Odyssey*, 2014, pp. 265–272.

[22] Zhengyang Chen, Shuai Wang, and Yanmin Qian, "Self-supervised learning based domain adaptation for robust speaker verification," in *ICASSP*. IEEE, 2021, pp. 5834–5838.

[23] Christian Szegedy et al., "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[24] Geoffrey Hinton et al., "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[25] Xun Huang and Serge Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of ICCV*, 2017, pp. 1501–1510.

[26] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *Proceedings of ICCV*, 2017, pp. 6924–6932.

[27] Xingang Pan, Ping Luo, et al., "Two at once: Enhancing learning and generalization capacities via IBN-NET," in *Proceedings of ECCV*, 2018, pp. 464–479.

[28] Xin Jin, Cuiling Lan, et al., "Style normalization and restitution for generalizable person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3143–3152.

[29] Jiyue Wang, Pei Zhang, Qianhua He, et al., "Revisiting label smoothing regularization with knowledge distillation," *Applied Sciences*, p. 4699, 2021.

[30] Li Yuan, Francis EH Tay, et al., "Revisiting knowledge distillation via label smoothing regularization," in *Proc. of CVPR*, 2020, pp. 3903–3911.

[31] Ying Liu, Yan Song, Yiheng Jiang, Ian McLoughlin, Lin Liu, and Lirong Dai, "An effective speaker recognition method based on joint identification and verification supervisions," in *INTERSPEECH*, 2020, pp. 3007–3011.