Interspeech 2024
1-5 September 2024, Kos, Greece

W-GVKT: Within-Global-View Knowledge Transfer for Speaker Verification

Zezhong Jin, Youzhi Tu, and Man-Wai Mak

Dept. of Electrical and Electronic Engineering, The Hong Kong Polytechnic University
Hong Kong SAR

zezhong.jin@connect.polyu.hk, 918tyz@gmail.com, enmwmak@polyu.edu.hk

Abstract

Contrastive self-supervised learning has played an important role in speaker verification (SV). However, such approaches suffer from false-negative issues. To address this problem, we enhance the non-contrastive DINO framework by enabling knowledge transfer from the teacher network to the student network through diversified versions of global views and call the method Within-Global-View Knowledge Transfer (W-GVKT) DINO. We discovered that given the global view of the entire utterance, creating discrepancies in the student's output through applying spectral augmentation and feature diversification to the global view can facilitate the transfer of knowledge from the teacher to the student. With negligible computational resource increases, W-GVKT achieves an impressive EER of 4.11% without utilizing speaker labels on Voxceleb1. When combined with the RDNIO framework, W-GVKT achieved an EER of 2.89%.

Index Terms: speaker verification, self-supervised learning, knowledge transfer, DINO

1. Introduction

In recent years, deep learning has experienced rapid development and has been widely applied in various domains. However, in most cases, the learning is supervised, which relies heavily on a large amount of labeled data for model training. Acquiring such labeled data is resource intensive. The possibility of dispensing with labeled data in self-supervised learning provides a perfect solution to the data labeling issues in many applications, such as computer vision [1–4], natural language processing [5, 6], and speaker verification (SV) [7–17].

Self-supervised learning can be divided into two categories: contrastive and non-contrastive. In SV, self-supervised contrastive learning aims to reduce the distance between (positive) samples from the same speaker and maximize the distance between (negative) samples from different speakers. Noting that separating channel and speaker information is vital in SV, the authors of [18] combined adversarial training with self-supervised contrastive training to prevent the speaker embedding network from learning channel information. The MoCo framework was applied to speaker verification (SV) in [7], alleviating the burden of requiring large batch size. The authors in [19] incorporated DSVAE into the SimCLR and MoCo frameworks, aiming to disentangle speaker information from non-speaker information in audio.

The aforementioned approaches assume that all samples in a mini-batch come from different speakers. However, in reality, we do not have the access to speaker labels and cannot guarantee this condition. In contrastive learning, for each anchor utterance in a mini-batch, utterances spoken by other speakers in the mini-batch are considered negatives. However, if a negative utterance shares the same speaker identity as the anchor, this false-negative utterance will cause contrastive learning to push the embeddings of the negative utterance and the anchor utterance apart. This issue, known as the class collision problem, can significantly impact the performance of self-supervised contrastive learning [20].

Many researchers have shifted their focus to non-contrastive frameworks to address the issue of false negatives. For example, researchers in computer vision have proposed a novel method called BYOL (Bootstrap Your Own Latent) [21] by only considering positive sample pairs in self-supervised learning, avoiding the problem of false negatives. Leveraging the success of BYOL, a subsequent study developed a knowledge distillation method with no labels (DINO) [4], streamlining BYOL's model architecture with a more efficient training strategy.

DINO is a self-distillation framework that consists of two components: a student network and a teacher network. Unlike traditional distillation methods, the student and teacher networks have the same structure but different parameters. For SV tasks, an utterance is cropped and segmented into global and local views of different durations (the long segments are referred to as the global view, while the short segments are referred to as the local views). After applying data augmentation, both the global and local views are fed into the student network, while only the global views are inputted into the teacher network. The learning process involves minimizing the cross-entropy between the output distributions of the student and teacher networks. The teacher network's parameters are updated by the exponential moving average (EMA) algorithm [3].

Several researchers have applied DINO to SV. For instance, a clustering approach was utilized to obtain more reasonable global and local views for DINO in [22]. The effectiveness of curriculum learning in the DINO framework was demonstrated in [12]. The authors in [23] introduced two regularization terms to address the issue of model collapse in DINO. However, none of the aforementioned methods have paid attention to how to enhance the knowledge transfer from the teacher network to the student network, which we believe is crucial for the success of the DINO framework. Based on this argument, we propose a simple yet effective framework to enhance the knowledge transfer from the teacher to the student.

Knowledge distillation in DINO is achieved by minimizing the cross-entropy between the output distributions of the teacher and student networks using different audio segments of an utterance as input. DINO excludes the cross-entropy between the teacher network's output and the student network's output arising from the same global view. However, we advo-

This work was supported by the RGC of Hong Kong SAR, Grant No. PolyU 15210122

10.21437/Interspeech.2024-354

cate that transferring knowledge from the teacher to the student based on the same global view can also benefit the student. Because our method allows knowledge transfer within the same global view, we call it "**w**ithin-**g**lobal-**v**iew **k**nowledge **t**ransfer (W-GVKT)" to differentiate it from the conventional DINO. We also introduced two strategies, spectral augmentation enhancement (SAE) and complementary feature enhancement (CFE), to increase the information diversity between the teacher and student networks within a global view. We are the first to leverage the knowledge distillation through diversifying the global views in SV tasks. Our contribution is that the proposed within-global-view knowledge transfer enables the student to receive the teacher's knowledge through diversified global views. The transfer is achieved without adding complex modules to the DINO framework, achieving significant performance gain with negligible computation overhead.

2. Methods

2.1. DINO for Speaker Verification

DINO [4] transfers knowledge from a teacher to a student through the cross-entropy loss at their outputs and the augmented views of objects at their inputs. In the context of SV, the student network comprises an encoder (f_s) that processes frame-level features and a projection network (h_s) that processes segment-level features, as shown in the lower branch of Figure 1. The teacher's encoder (f_t) and projection network (h_t) have the same structure as the student network but with different parameters.

We randomly sample one long segment $\mathbf{x}_g$ and two short segments $\{\mathbf{x}_{l1}, \mathbf{x}_{l2}\}$ from one utterance, where g and l indicate that the corresponding segments belong to the global and local views, respectively. After augmentation and feature extraction, we obtain filter-bank (FBank) features $\boldsymbol{X}_g$, $\boldsymbol{X}_{l1}$, and $\boldsymbol{X}_{l2}$, respectively (see the "Random Cropping & Augmentation block" in Figure 1). We feed $\mathcal{X}^t = \{\boldsymbol{X}_g\}$ into the teacher network and $\mathcal{X}^s = \{\boldsymbol{X}_g, \boldsymbol{X}_{l1}, \boldsymbol{X}_{l2}\}$ into the student network. We obtain probability vectors $\mathcal{P}^t = \{\boldsymbol{y}^t_{\boldsymbol{X}_g}\}$ and $\mathcal{P}^s = \{\boldsymbol{y}^s_{\boldsymbol{X}_g}, \boldsymbol{y}^s_{\boldsymbol{X}_{l1}}, \boldsymbol{y}^s_{\boldsymbol{X}_{l2}}\}$ at the softmax layer of the projection network (MLP), where $\boldsymbol{y}$ is a vector of K dimensions. To prevent model collapse, the output of the teacher's projection head undergoes a centering operation before the softmax function.

The student network is optimized by minimizing the cross-entropy between the distributions of the student's and teacher's outputs. The loss for each utterance is defined as:

$$L_{\text{DINO}} = \sum_{\boldsymbol{X} \in \mathcal{X}^t} \sum_{\substack{\boldsymbol{X}' \in \mathcal{X}^s \\ \boldsymbol{X}' \neq \boldsymbol{X}}} \text{CrossEntropy}(\boldsymbol{y}^t_{\boldsymbol{X}}, \boldsymbol{y}^s_{\boldsymbol{X}'}), \tag{1}$$

where $\text{CrossEntropy}(\boldsymbol{a}, \boldsymbol{b}) = -\sum_{k=1}^{K} a_k \log b_k$. The teacher network ($f_t$ and h_t) are updated from the student network (f_s and h_s) using the exponential moving average (EMA) algorithm [3]. After training, the vectors outputted from the teacher encoder f_t are used as speaker embeddings.

2.2. W-GVKT DINO

Due to the constraint $\boldsymbol{X}' \neq \boldsymbol{X}$ in Eq. 1, DINO does not compare the teacher's and student's outputs of the same global segment $\mathbf{x}_g$. Because every term in Eq. 1 represents a certain knowledge transfer, it will be beneficial to the knowledge transfer by diversifying the input instead of omitting some terms. We propose diversifying the input by creating multiple augmented views of the entire utterance and including these augmented global views in the cross-entropy computation. We also propose using spectral augmentation [24] and complementary features to enhance the diversity of the augmented views. They are named SpecAugment Enhancement (SAE) and Complementary Feature Enhancement (SFE). The diversification is implemented by the "Random Cropping & Augmentation" block in Figure 1.

For SAE, we apply SpecAugment [24] to $\boldsymbol{X}_g$ to create $\boldsymbol{X}_{g\text{-SpecAug}}$ by masking a portion of the frequency bins of the Fbank features at certain time intervals. For CFE, we extract different acoustic features, such as mel-frequency cepstral coefficients (MFCCs), linear frequency cepstral coefficients (LFCCs), and a spectrogram from $\mathbf{x}_g$ to create $\boldsymbol{X}_{g\text{-MFCC}}$, $\boldsymbol{X}_{g\text{-LFCC}}$, and $\boldsymbol{X}_{g\text{-Spectro}}$, respectively. Taking LFCCs as an example, we obtain $\mathcal{X}^t = \{\boldsymbol{X}_g\}$ and $\mathcal{X}^s = \{\boldsymbol{X}_{g\text{-SpecAug}}, \boldsymbol{X}_{g\text{-LFCC}}, \boldsymbol{X}_{l1}, \boldsymbol{X}_{l2}\}$ as the input to the teacher and student networks, respectively. The W-GVKT DINO loss for each utterance is defined as:

$$L_{\text{W-GVKT-DINO}} = \sum_{\boldsymbol{X} \in \mathcal{X}^t} \sum_{\boldsymbol{X}' \in \mathcal{X}^s} \text{CrossEntropy}(\boldsymbol{y}^t_{\boldsymbol{X}}, \boldsymbol{y}^s_{\boldsymbol{X}'}). \tag{2}$$

As observed from [22,23], the diversity of input data or embeddings is important to prevent DINO from falling into trivial solutions. In W-GVKT DINO, we not only focus on how to enhance knowledge distillation from the teacher to the student but also use SAE and CFE to increase input information diversity to prevent the system from falling into trivial solutions. By applying the augmented versions and various acoustic features of the same global view, diverse information can be distilled from the teacher to the student.

3. Experimental Setup

3.1. Datasets

We used Voxceleb2 [25] as the training set. This dataset comprises 1,092,009 utterances from 5,994 speakers. During the training process, we did not use the speaker labels. The performance of the systems was evaluated on Vox1-O, Vox1-E, and Vox1-H [26]. We followed the Kaldi recipe for data augmentation, incorporating noise from the MUSAN [27] dataset and reverberation using the RIR [28] dataset.

3.2. System Configuration

We utilized the small ECAPA-TDNN [29] with 512 channels as the backbone. The MLP in Figure 1 contains three fully connected layers, with a hidden layer of 2048 dimensions, followed by an L_2-normalization layer and a weight normalization layer [30]. The MLP maps the speaker embeddings to an output layer with K dimensions. We set K to 65,536, the same as the configuration in [4]. Although this value is much larger than the actual number of speakers (5,994) in Voxceleb2, the DINO and W-GVKT DINO work well under this setting. A reason for the good performance is that knowledge transfer could occur as long as the hypothetic "classes" are distinguishable. It does not matter if each class corresponds to a speaker. In our case, each speaker may belong to many hypothetic classes because the durations of the utterances are rather short in our experiments.

In W-GVKT DINO, we chose three different hand-crafted features, LFCCs, MFCCs, and spectrograms, as the global views of an utterance. The LFCCs and MFCCs have a dimension of 40, while the spectrograms have a dimension of 257.

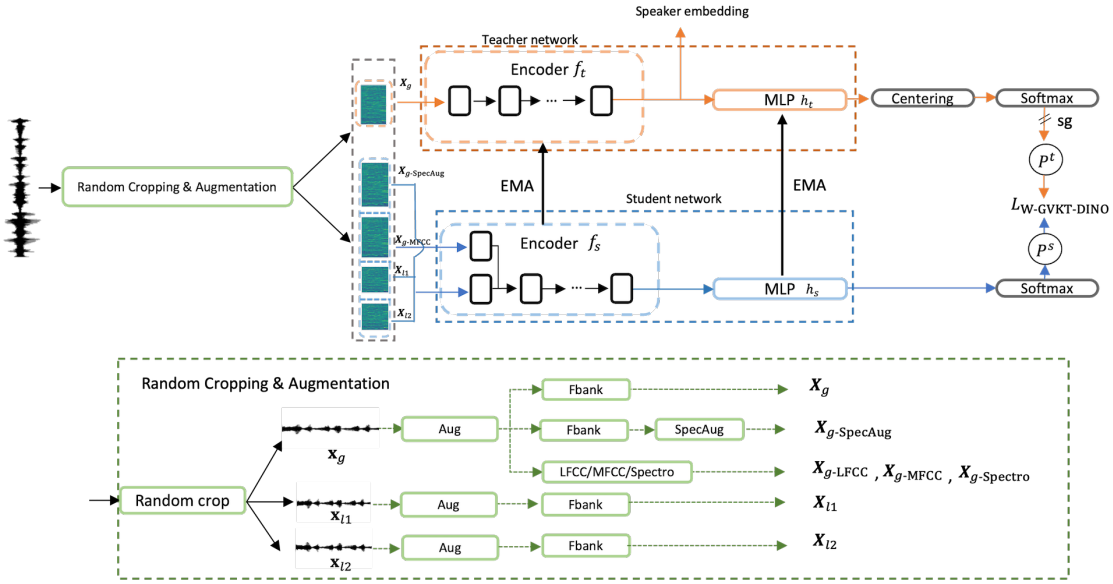

Figure 1: *The Framework of W-GVKT DINO. EMA and sg stand for exponential moving average and stop gradient, respectively. The first layer of the student network is divided into two branches to handle features of different dimensions. The lower branch of the student encoder is used for the EMA update of the teacher encoder.*

Because the FBank, MFCC (or LFCC), and spectrogram features have different dimensions, the convolutional operations in the first layer of the student network were divided into two branches, as shown in Figure 1. The parameters of the teacher encoder were updated by EMA using the lower branch of the student encoder. The teacher's MLP was updated by EMA using the student's MLP. Therefore, the part of the student network used for the EMA update has the same structure as the teacher network.

Each global view was derived from a segment with a duration of 3 seconds, while the segments for deriving the local views were limited to 2 seconds. The temperature parameter of the student network's softmax function was set to 0.1. The temperature parameter for the teacher's softmax function was linearly increased from 0.04 to 0.07 during the first 30 epochs and fixed afterward. For SpecAugment, we set the frequency masking parameter to 6 and the time masking parameter to 10. We used the SGD optimizer and the Cosine scheduler [21] to optimize the models. The initial learning rate was set to 0.2, and the final learning rate was set to 0.00005.

Results are presented using two metrics: equal error rate (EER) and minimum detection cost function (MinDCF). The MinDCF was computed with the settings $P_{\text{target}} = 0.01$ and $C_{\text{fa}} = C_{\text{miss}} = 1$. The cosine similarity was employed to calculate the scores.

4. Results and Disscussions

4.1. Main Results

Table 1 presents the main experimental results. The first row corresponds to the baseline using Eq. 1 as the loss function, i.e., there is no knowledge transfer from the teacher to the student through the same global view. The second row corresponds to the scenario where knowledge can be transferred through the same global view. However, the two strategies (SAE and CFE) for enhancing the information diversity among the same global view were not used. The third row represents the case where $\boldsymbol{X}_{g\text{-SpecAug}}$ was fed into the student network instead of $\boldsymbol{X}_g$. The fourth row corresponds to the scenario where both $\boldsymbol{X}_{g\text{-SpecAug}}$ and $\boldsymbol{X}_{g\text{-LFCC}}$ were fed into the student network. We have reproduced the work of RDINO [23] and conducted experiments using W-GVKT RDINO. The results are shown in Rows 5 and 6.

We conducted experiments on three test scenarios in VoxCeleb1, and our method (W-GVKT DINO) consistently outperforms the baseline across all scenarios (improvement of 19% on Vox1-O, 17% on Vox1-E, and 17% on Vox1-H). These results demonstrate the effectiveness of W-GVKT DINO. Comparing the first and second rows of Table 1, we observe that distilling knowledge through diversified global view can indeed improve performance. Rows 3 and 4 show that applying SAE and CFE on the same global view can enhance the effectiveness of knowledge transfer from the teacher to the student. Rows 5 and 6 demonstrate the effectiveness of combining W-GVKT with RDINO. It shows that under the standard DINO setting (2 globals and 4 locals), W-GVKT remains effective and achieves competitive performance.

To further verify the benefits of knowledge transfer via the diversified global views, we compared W-GVTK DINO with other recent self-supervised methods. Because the model size of RDINO in [23] differs from ours, readers should not directly compare the performance of RDINO in [23] with the performance in Tables 1 and 2. To have a fair comparison, we re-implemented RDINO using the ECAPA-TDNN as the backbone. Table 2 shows that W-GVKT DINO outperforms the previous self-supervised methods. Notably, the systems in [9, 12, 14] utilize a more computationally expensive setup with 2 global and 4 local views. Additionally, the length of the global views in these studies was set to 4 seconds, whereas ours was set to 3 seconds only. Therefore, our method achieves superior performance while conserving training resources.

Table 1: *Comparison of the Proposed W-GVKT DINO with the baseline on Voxceleb1 test sets. In RDINO, we used a small ECAPA-TDNN with 512 channels. The other experimental settings were kept the same as [23].*

Row	System	Vox1-O		Vox1-E		Vox1-H	
		EER (%)	minDCF	EER (%)	minDCF	EER (%)	minDCF
1	DINO (Baseline)	5.07	0.582	5.93	0.682	10.82	0.839
2	DINO+W-GVKT (no SAE and CFE)	4.80	0.522	5.67	0.592	10.32	0.766
3	DINO+W-GVKT (w/ SAE)	4.76	0.513	5.61	0.567	9.70	**0.720**
4	DINO+W-GVKT (w/ SAE & CFE)	**4.11**	**0.466**	**4.93**	**0.536**	**8.96**	0.750
5	RDINO [23]	3.85	0.401	4.26	0.493	7.24	0.689
6	RDINO+W-GVKT (SpecAug + LFCC)	**2.89**	**0.333**	**3.02**	**0.392**	**6.04**	**0.567**

Table 2: *Comparison of our method (W-GVKT) with other state-of-art methods on the Vox1-O test set. The minDCF with a * was calculated using Ptarget = 0.05 instead of Ptarget = 0.01.*

System	EER (%)	minDCF
AP+AAT [18]	8.65	0.454*
MoCo+WavAug [7]	8.23	0.590
Contrastive [10]	7.36	N/R
SSReg [13]	6.99	0.434*
DINO [9]	4.83	N/R
DINO+CL [12]	4.47	0.3057*
DINO+W-GVKT (Ours)	**4.11**	0.466
RDINO	3.85	0.401
RDINO+W-GVKT (Ours)	**2.89**	**0.333**

Table 3: *The Impact of applying CFE on different handcrafted features on the performance in Vox1-O.*

Handcrafted Feature	EER (%)	minDCF
MFCC	4.44	0.471
Spectrogram	4.51	0.485
LFCC	**4.11**	**0.466**

4.2. Impact of CFE and SAE

We conducted experiments to investigate the effect of feature diversification through CFE and SAE on knowledge transfer. Table 3 presents the impact of using different handcrafted features for diversifying the global views. Results show that setting $\mathcal{X}^s = \{\boldsymbol{X}_{g\text{-SpecAug}}, \boldsymbol{X}_{g\text{-LFCC}}, \boldsymbol{X}_{l1}, \boldsymbol{X}_{l2}\}$ for the student network achieves the best performance.

We also conducted a series of experiments to explore the impact of time masking and frequency masking in SpecAugment, which introduces diversity in the same global view. Table 4 shows that the performance is the best when both time and frequency masks are applied.

4.3. Fine-tuning on Self-supervised Models

To further illustrate the effectiveness of W-GVKT, we investigate the performance of fine-tuning the self-supervised models with a small amount of labeled data in VoxCeleb1. Specifically, a classification head was added to the teacher encoder of the DINO variants, and the AAMSoftmax [31] loss was adopted to fine-tune the teacher encoder using the labeled data in the development set of Voxceleb1. As a comparison, we also trained a randomly initialized speaker encoder, whose performance is shown in the first row of Table 5.

Table 4: *Impact of masking time-, frequency, and time-frequency domains in SpecAugment on the performance in Vox1-O.*

Domain	EER (%)	minDCF
None	4.80	0.522
Time	4.81	0.533
Freq	4.87	0.561
Time-Freq	**4.76**	**0.513**

Table 5: *Performance of training a speaker encoder from scratch (Row 1) or fine-tuning variant of DINO-based self-supervised models (Rows 2–4) on the Vox1-O test set using the labeled data in the development set of Voxceleb1.*

Row	Initial Teacher Encoder	EER (%)	minDCF
1	Randomly Initialized	2.67	0.273
2	DINO (Baseline)	2.41	0.257
3	DINO+W-GVKT	2.16	0.242
4	RDINO+W-GVKT	**1.86**	**0.229**

Table 5 shows that fine-tuning the DINO encoder (Rows 2–4) achieves much better performance than training a speaker encoder from scratch (Row 1). With a small amount of labeled data, fine-tuning RDINO+W-GVKT leads to a 30.34% reduction in EER. Moreover, DINO with W-GVKT achieves much better performance than pure DINO, demonstrating that W-GVKT is effective for the DINO to find a good initial condition for supervised fine-tuning.

5. Conclusions

This paper investigates the knowledge distillation in DINO using the diversified global views of an utterance for speaker verification. The idea leads to an enhanced DINO called within-global-view knowledge transfer (W-GVKT) DINO, where knowledge distillation is promoted by using the augmented versions of the global views and diverse acoustic features derived from them. We demonstrate that this idea of global view diversification can boost the performance of DINO and the more recent RDINO framework for speaker verification.

6. References

[1] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. International Conference on Machine Learning*, 2020, pp. 1597–1607.

[3] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[4] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9650–9660.

[5] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[6] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger *et al.*, "ProtTrans: Toward understanding the language of life through self-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 7112–7127, 2021.

[7] W. Xia, C. Zhang, C. Weng, M. Yu, and D. Yu, "Self-supervised text-independent speaker verification using prototypical momentum contrastive learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6723–6727.

[8] D. Cai, W. Wang, and M. Li, "An iterative framework for self-supervised deep speaker representation learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6728–6732.

[9] J. Cho, J. Villalba, and N. Dehak, "The JHU submission to VoxSrc-21: Track 3," *arXiv preprint arXiv:2109.13425*, 2021.

[10] R. Tao, K. A. Lee, R. K. Das, V. Hautamäki, and H. Li, "Self-supervised speaker recognition with loss-gated learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6142–6146.

[11] J.-W. Jung, Y. J. Kim, H.-S. Heo, B.-J. Lee, Y. Kwon, and J. S. Chung, "Pushing the limits of raw waveform speaker recognition," *in Proc. Interspeech*, pp. 2228–2232, 2022.

[12] H.-S. Heo, J.-W. Jung, J. Kang, Y. Kwon, Y. J. Kim, and B.-J. L. JS Chung, "Self-supervised curriculum learning for speaker verification," *arXiv preprint arXiv:2203.14525*, 2022.

[13] M. Sang, H. Li, F. Liu, A. O. Arnold, and L. Wan, "Self-supervised speaker verification with simple siamese network and self-supervised regularization," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6127–6131.

[14] B. Han, Z. Chen, and Y. Qian, "Self-supervised speaker verification using dynamic loss-gate and label correction," *in Proc. Interspeech*, pp. 4780–4784, 2022.

[15] S. Zheng, G. Liu, H. Suo, and Y. Lei, "Autoencoder-based semi-supervised curriculum learning for out-of-domain speaker verification," *in Proc. Interspeech*, pp. 1440–1444, 2019.

[16] ——, "Towards a fault-tolerant speaker verification system: A regularization approach to reduce the condition number." *in Proc. Interspeech*, pp. 4065–4069, 2019.

[17] H. Mao, F. Hong, and M.-W. Mak, "Cluster-guided unsupervised domain adaptation for deep speaker embedding," *IEEE Signal Processing Letters*, vol. 30, pp. 643–647, 2023.

[18] J. Huh, H. S. Heo, J. Kang, S. Watanabe, and J. Son Chung, "Augmentation adversarial training for self-supervised speaker recognition," in *Proc. Workshop on Self-Supervised Learning for Speech and Audio Processing, NeurIPS*, 2020.

[19] Y. Tu, M.-W. Mak, and J.-T. Chien, "Contrastive self-supervised speaker embedding with sequential disentanglement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2704–2715, 2024.

[20] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[21] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent: a new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 271–21 284, 2020.

[22] B. Han, Z. Chen, and Y. Qian, "Self-supervised learning with cluster-aware-DINO for high-performance robust speaker verification," *arXiv preprint arXiv:2304.05754*, 2023.

[23] Y. Chen, S. Zheng, H. Wang, L. Cheng, and Q. Chen, "Pushing the limits of self-supervised speaker verification using regularized distillation framework," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[24] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *in Proc. Interspeech*, pp. 2613–2617, 2019.

[25] J. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *in Proc. Interspeech*, pp. 1086–1090, 2018.

[26] A. Nagrani, J. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *in Proc. Interspeech*, pp. 2616—-2620, 2017.

[27] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[28] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[29] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," *in Proc. Interspeech*, pp. 3830–3834, 2020.

[30] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," *Advances in Neural Information Processing Systems*, vol. 29, pp. 458–463, 2016.

[31] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 1652–1656.