



# Pushing the Frontiers of Self-Distillation Prototypes Network with Dimension Regularization and Score Normalization

Yafeng Chen<sup>1</sup>, Chong Deng<sup>1</sup>, Hui Wang<sup>1</sup>, Yiheng Jiang<sup>1</sup>, Han Yin<sup>1</sup>, Qian Chen<sup>1</sup>, Wen Wang<sup>1</sup>

<sup>1</sup>Tongyi Speech Lab, Alibaba Group, China

chenyafeng.cyf@alibaba-inc.com

## Abstract

Developing robust speaker verification (SV) systems without speaker labels has been a longstanding challenge. Earlier research has highlighted a considerable performance gap between self-supervised and fully supervised approaches. In this paper, we enhance the non-contrastive self-supervised framework, Self-Distillation Prototypes Network (SDPN), by introducing dimension regularization that explicitly addresses the collapse problem through the application of regularization terms to speaker embeddings. Moreover, we integrate score normalization techniques from fully supervised SV to further bridge the gap toward supervised verification performance. SDPN with dimension regularization and score normalization sets a new state-of-the-art on the VoxCeleb1 speaker verification evaluation benchmark, achieving Equal Error Rate **1.29%**, **1.60%**, and **2.80%** for trial VoxCeleb1-{O,E,H} respectively.<sup>1</sup> These results demonstrate relative improvements of **28.3%**, **19.6%**, and **22.6%** over the current best self-supervised methods, thereby advancing the frontiers of SV technology.

**Index Terms:** speaker verification, self-distillation prototypes network, dimension regularization, score normalization

## 1. Introduction

Deep learning methods have significantly advanced the performance of speaker verification (SV) tasks. These improvements have been driven by the availability of large labeled datasets and effective data augmentation methods. However, collecting extensive labeled data in the real world is both time-consuming and costly. As an alternative, self-supervised learning (SSL), which relies on unlabeled data, offers a promising solution for learning robust speech representations. Self-supervised methods are typically categorized into three main directions: contrastive [1–6], non-contrastive [7–12], and dimension contrastive [13–15] learning.

Contrastive learning frameworks [1–6], such as SimCLR [1] and MoCo [2], establish instance-level discriminative capabilities through positive/negative pair optimization. While effective in learning representations that are invariant to data augmentation, these methods face challenges due to their reliance on large batch sizes (SimCLR) or complex memory mechanisms (MoCo), which lead to computational bottlenecks. Xia et al. [3] introduce SimCLR and MoCo frameworks with designed augmentation strategies for self-supervised speaker verification, achieving performance improvements.

Non-contrastive methods [7–12] avoid negative pairs, introducing the collapse problem where speaker representations

converge to trivial solutions. BYOL [7] uses asymmetric architectures with predictors and stop-gradient mechanisms to prevent collapse, while DINO [8] leverages knowledge distillation between teacher-student networks to maintain consistency. Inspired by BYOL, Sang et al. [9] propose a self-supervised SV framework that focuses on the similarity between the latent representations of positive data pairs. Zhang et al. [10] improve SV performance by applying different augmentation strategies to DINO. Chen et al. [12] propose the Self-Distillation Prototypes Network (SDPN) to capture the relationship between representations of utterances from different speakers by integrating learnable prototypes into a self-distillation framework. SDPN achieves excellent performance with low computational resources, comparing favorably to many DINO-based methods.

Dimension-contrastive methods [13–15] introduce a novel paradigm by optimizing cross-dimension relationships rather than instance discrimination. Barlow Twins [13] minimizes correlation between dimensions to enforce independence, while VICReg [14] learns invariance to different views, avoiding collapse of the representations with a variance preservation term, and maximizing the information content of the representation with a covariance regularization term. Building on these insights, Garrido et al. [15] demonstrate the complementary nature of sample-contrastive (including both contrastive and non-contrastive paradigms) and dimension-contrastive approaches.

Inspired by [14, 15], we optimize SDPN by introducing two dimension regularization terms. These additions aim to reduce the risk of collapse and improve the robustness of speaker verification by minimizing the correlation between different embedding dimensions and increasing the diversity of feature dimensions. The first regularization term is off-diagonal dimension regularization, and the second is Frobenius dimension regularization. Both techniques work to decorrelate the variables of each embedding while minimizing redundancy.

Score normalization is a common post-processing step used to align scores from different trials to a common scale. It requires no additional parameters and incurs minimal computational overhead. Most self-supervised speaker verification methods [9, 10, 12, 16–18] employ cosine similarity for scoring. However, score distributions often vary due to factors such as channel, duration, and gender differences between datasets. A fixed threshold can adversely affect overall verification performance and exacerbate the score drift phenomenon, which is particularly pronounced in self-supervised learning.

To address this issue, we integrate normalization algorithms from supervised frameworks into SDPN with dimension regularization. This approach achieves new state-of-the-art results on VoxCeleb1, with equal error rates of 1.29%, 1.60%, and 2.80% for the VoxCeleb1-O, VoxCeleb1-E, and VoxCeleb1-H trials, without using any speaker labels during training.

<sup>1</sup>Code will be publicly available at <https://github.com/modelscope/3D-Speaker>

## 2. Preliminary Knowledge

### 2.1. Self-distillation prototypes network

The self-distillation prototypes network [12] integrates a teacher and a student network, both sharing the same architecture but differing in parameters. As a non-contrastive self-supervised framework, it uses the teacher network’s outputs as targets to optimize the student network concurrently. Each network comprises three primary components: an encoder for extracting speaker embeddings, a multi-layer perceptron for feature transformation, and learnable prototypes to compute soft distributions between global and local views. Here, global views  $\mathbf{X}_g = \{\mathbf{x}_g\}$  are long segments, while local views  $\mathbf{X}_l = \{\mathbf{x}_{l_1}, \mathbf{x}_{l_2}, \mathbf{x}_{l_3}, \mathbf{x}_{l_4}\}$  are short segments, both randomly derived from the same utterance. Consider the teacher module as an example: it consists of a backbone  $f$  and a projection head  $h$ . The speaker embedding is derived from the output of the backbone  $f$ . The projection head  $h$  includes three fully connected layers with dimensions 2048-2048-256, followed by  $L2$  normalization. The learnable prototypes  $\mathbf{C}$  are shared between the teacher and student networks and are used to compute the soft distributions between global and local views. The cross-entropy loss is then calculated to minimize the probability distribution as follows:

$$\mathcal{L}_{CE} = \sum_{\mathbf{x} \in \mathbf{X}_g} \sum_{\mathbf{x}' \in \mathbf{X}_l} H(P^{tea}(\mathbf{x}) | P^{stu}(\mathbf{x}')) \quad (1)$$

where  $H(a|b) = -a \log b$  is cross-entropy.  $P^{tea}$  and  $P^{stu}$  denote the output probability distributions of the teacher network and the student network.

To prevent the collapse problem, SDPN introduces a diversity regularization term. This approach assesses the pairwise similarity among embeddings ( $\mathbf{x}_u$  and  $\mathbf{x}_v$ ), deliberately separating the closest embeddings to enhance the diversity of speaker embeddings within a batch, thereby enriching the learning process and enhancing the model’s generalization capabilities. The diversity regularization loss is calculated as follows:

$$\mathcal{L}_{RE} = -\frac{1}{n} \sum_{u=1}^n \left( \sum_{v=1, v \neq u}^n \log(\min_{v \neq u} \|\mathbf{x}_u - \mathbf{x}_v\|) \right) \quad (2)$$

where  $n$  is the batch size.

The overall training objective of SDPN is to minimize a combination of the CE loss and diversity regularization loss, weighted by the hyperparameter  $\mu$ .

$$\mathcal{L}_{SDPN} = \mathcal{L}_{CE} + \mu \mathcal{L}_{RE} \quad (3)$$

### 2.2. Score normalization in supervised SV

The score distribution of the test set is typically unknown in advance. Researchers propose various score normalization methods [19–21] to minimize the discrepancy between estimated and true distributions, such as zero normalization (Z-norm) [19], test normalization (T-norm) [19], symmetric normalization (S-norm) [20], and adaptive symmetric score normalization (AS-norm) [21]. Z-norm normalizes the score distribution of the target speaker model. T-norm normalizes the score distribution of the impostors. S-norm combines the benefits of Z-norm and T-norm by averaging their normalized scores, ensuring that the score remains consistent. Furthermore, AS-norm introduces adaptive cohort selection, choosing the top  $K$  scores to calculate normalization parameters, thus aligning the estimated score distribution more closely with the actual test set distribution.

It involves the use of a cohort  $U = \{u_i\}_{i=1}^N$  consisting of  $N$  speakers.  $S_e$  are obtained by scoring the enrollment utterance  $e$  against all files from the cohort  $U$ .  $S_t$  are obtained by scoring the test utterance  $t$  against all files from the cohort  $U$ . Specifically, the AS-norm formula is given by:

$$s(e, t)_{as-norm} = \frac{1}{2} \cdot \left( \frac{s(e, t) - \mu(S_e(U_e^{top}))}{\sigma(S_e(U_e^{top}))} + \frac{s(e, t) - \mu(S_t(U_t^{top}))}{\sigma(S_t(U_t^{top}))} \right) \quad (4)$$

where  $s(e, t)$  is the original score between the enrollment utterance  $e$  and the test utterance  $t$ .  $U_e^{top}$  and  $U_t^{top}$  represent the selected top  $K$  cohort members for the enrollment and test utterances, respectively.  $\mu(S_e(U_e^{top}))$  and  $\sigma(S_e(U_e^{top}))$  denote the mean and standard deviation of the top  $K$  cohort scores  $S_e$  for the enrollment utterance  $e$ .  $\mu(S_t(U_t^{top}))$  and  $\sigma(S_t(U_t^{top}))$  are the mean and standard deviation of the top  $K$  cohort scores  $S_t$  for the test utterance  $t$ .

## 3. Proposed Method

To further alleviate the collapse problem within the SDPN framework, we introduce two dimensional regularization terms: off-diagonal dimension regularization and Frobenius dimension regularization. Additionally, we boost the SDPN with dimension regularization by incorporating various score normalization algorithms, which significantly improve the performance of self-supervised speaker verification.

### 3.1. Off-diagonal dimension regularization

The off-diagonal dimension regularization term is designed to decorrelate the variables within each embedding. By driving the covariances of each dimension in all embeddings within a batch towards zero, this approach prevents informational collapse, where variables become highly correlated. The off-diagonal dimension regularization loss is computed as follows:

$$L_{ODR} = \sum_i^d \sum_{j \neq i}^d C_{ij}^{tea^2} + \sum_i^d \sum_{j \neq i}^d C_{ij}^{stu^2} \quad (5)$$

where  $C$  is the covariance matrix computed from the global outputs of teacher network and student network along the batch dimension, and  $d$  is matrix dimension.  $C_{ij}$  is defined as Eq. 6:

$$C_{ij} = \frac{\sum_b z_{b,i} z_{b,j}}{\sqrt{\sum_b (z_{b,i})^2} \sqrt{\sum_b (z_{b,j})^2}} \quad (6)$$

where  $b$  indexes batch samples and  $i, j$  index the embedding dimension. This term encourages the off-diagonal coefficients of  $C_{ij}$  to be close to 0, decorrelating the different dimensions of the embeddings and preventing them from encoding similar information.

### 3.2. Frobenius dimension regularization

The Frobenius dimension regularization term minimizes the correlation between different embedding dimensions, aiming to reduce redundancy and enhance feature diversity across dimensions. It is calculated as the logarithm of the squared Frobenius norm of the normalized covariance embedding matrices, denoted as  $C$ . The Frobenius norm of  $C$  is defined as follows:

$$\|C\|_{Frob} = \sqrt{\sum_i^d \sum_j^d C_{ij}^2} \quad (7)$$

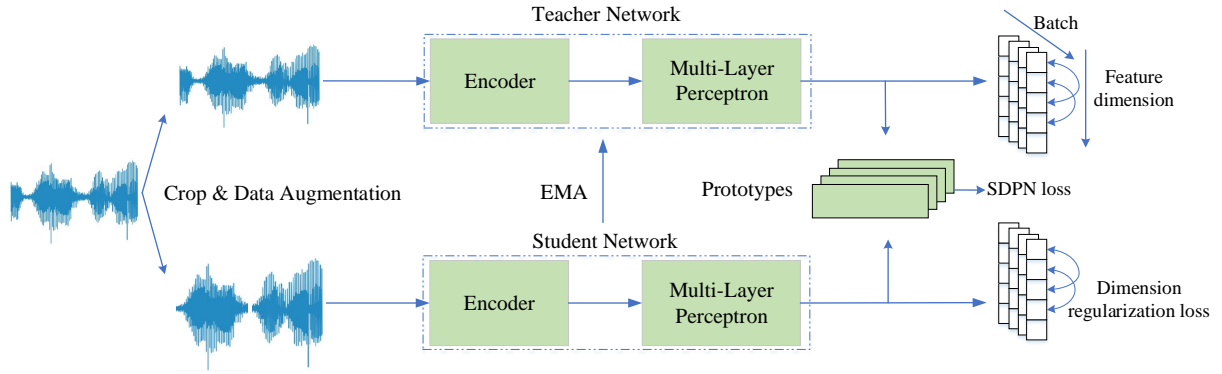


Figure 1: Overview of the SDPN framework with dimension regularization: It includes teacher and student networks with identical architectures but different parameters. The teacher network’s outputs serve as targets to optimize the student network. Diversity regularization reduces the correlation between feature dimensions.

$$L_{FDR} = \log(\|C\|_{Frob}^{tea}) + \log(\|C\|_{Frob}^{stu}) \quad (8)$$

The gradient of the  $L_{FDR}$  function can be formulated as:

$$\frac{\partial L_{FDR}}{\partial C_{ij}} = \begin{cases} \frac{C_{ij}}{D + \sum_{i \neq j} C_{ij}^2}, & \text{for } i \neq j \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

Different from the off-diagonal dimension regularization, this formulation exhibits two critical properties:

- **Gradient magnitude modulation:** The denominator  $D + \sum_{i \neq j} C_{ij}^2$  provides adaptive stabilization. When inter-dimensional correlations are weak, the gradients are bounded by the constant term  $D$ , preventing explosive updates. This ensures that the model’s learning dynamics remain stable without excessively large gradient updates, allowing for controlled and consistent learning.
- **Implicit annealing effect:** As training progresses and inter-dimensional correlations diminish, the denominator undergoes a transition from being dominated by  $D$  to being dominated by the correlations. This natural decay of the learning rate facilitates a more refined optimization process.

### 3.3. SDPN with dimension regularization

The overview of SDPN framework with dimension regularization is depicted in Fig. 1. The speaker embedding network is jointly trained with the  $L_{SDPN}$  and  $L_{DR}$  ( $L_{ODR}$  or  $L_{FDR}$ ). The overall loss is calculated as Eq. 10, the hyperparameter  $\lambda$  control the balance of losses.

$$\mathcal{L} = \mathcal{L}_{SDPN} + \lambda \mathcal{L}_{DR} \quad (10)$$

### 3.4. Score normalization in self-supervised framework

In the field of self-supervised speaker verification, limited research has addressed score normalization to tackle inconsistent score distribution, despite the more pronounced issue of score drifting compared to fully supervised scenarios.

Among the various self-supervised frameworks available, we choose the SDPN framework with dimension regularization for its outstanding performance to evaluate the effectiveness of several score normalization techniques. Z-norm and T-norm are fundamental approaches that standardize scores based on global

statistics, including the mean and standard deviation. AS-norm merges the benefits of Z-norm and T-norm. Its normalization parameters are calculated using the mean and standard deviation of the highest scores. This methodology aids in diminishing the discrepancy between the estimated score distribution and the actual distribution within the test set, thereby mitigating the loss of distributional information when employing a fixed quantity of score statistics as parameters.

## 4. Experiments and analysis

### 4.1. Experimental settings

#### 4.1.1. Datasets and evaluation metrics

To evaluate the effectiveness of the proposed method, we conduct experiments using the VoxCeleb datasets. The development portion of VoxCeleb2 [22], consisting of 1,092,009 utterances from 5,994 speakers, is utilized for training. The performance of all systems is assessed on the test set of VoxCeleb1 [23]. No speaker labels are used during training in any of the experiments. The results are presented using two metrics: the equal error rate (EER) and the minimum of the normalized detection cost function (MinDCF), with the parameters set to  $P_{target} = 0.05$  and  $C_{fa} = C_{miss} = 1$ .

#### 4.1.2. Input features

For each utterance, we employ a multi-crop strategy in SDPN training, utilizing 4-second segments as global views and 2-second segments as local views. The acoustic features in the experiments are 80-dimensional Filter Bank (FBank) with 25ms windows and a 10ms shift. Speech activity detection (SAD) is not applied, as the training data predominantly comprises continuous speech. Mean and variance normalization are conducted using instance normalization on the FBank features. WavAugment and SpecAugment are used in training process.

### 4.2. Model configurations and implementation details

We exploit the ECAPA-TDNN with attentive statistical pooling as the encoder  $f$ , followed by a 512-d FC layer. The projection head  $h$  consists of four FC layers with hidden size of 3072-3072-1024. We train the model 160 epochs using the stochastic gradient descent (SGD) optimizer with momentum of 0.9, on 8

Table 1: Results on VoxCeleb1-O, VoxCeleb1-E, and VoxCeleb1-H datasets. DINO\* and SDPN\* refers to our replication of the DINO and SDPN framework respectively. SDPN w/ off-diagonal denotes adding the off-diagonal dimension regularization to SDPN. SDPN w/ FroNorm denotes adding the Frobenius dimension regularization to SDPN. The best results for each test set are in bold.

Architecture	Score Normalization	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
		EER (%)↓	minDCF↓	EER (%)↓	minDCF↓	EER (%)↓	minDCF↓
DINO*	Cosine	2.65	0.202	2.74	0.188	5.02	0.304
SDPN*	Cosine	1.80	0.139	1.99	0.131	3.62	0.219
SDPN w/ off-diagonal	Cosine	1.69	0.128	1.89	0.123	3.43	0.208
	Z-norm	1.48	0.119	1.77	0.113	3.01	0.194
	T-norm	1.52	0.115	1.76	0.116	3.04	0.192
	AS-norm	1.39	0.102	1.71	0.103	2.87	0.176
SDPN w/ FroNorm	Cosine	1.63	0.124	1.86	0.121	3.38	0.203
	Z-norm	1.37	0.111	1.68	0.103	2.95	0.188
	T-norm	1.37	0.105	1.66	0.105	2.93	0.187
	AS-norm	<b>1.29</b>	<b>0.096</b>	<b>1.60</b>	<b>0.094</b>	<b>2.80</b>	<b>0.169</b>

Table 2: Comparison of the results of SDPN with FroNorm and AS-norm against those of recent SSL models on the VoxCeleb-O.

Model	Extractor	EER(%)↓
SSReg [9]	Fast ResNet34	6.99
MoCo-DSVAE [6]	ECAPA-TDNN	6.29
Mixup-Aug [4]	Fast ResNet34	5.84
DINO + CL [16]	ECAPA-TDNN	4.47
DINO [10]	ECAPA-TDNN	3.30
MeMo-CTES [17]	ECAPA-TDNN	3.10
PDC-DINO [18]	ECAPA-TDNN	3.05
C3-DINO [10]	ECAPA-TDNN	2.50
SDPN [12]	ECAPA-TDNN	1.80
<b>SDPN w/ FroNorm</b>	<b>ECAPA-TDNN</b>	<b>1.63</b>
<b>SDPN w/ FroNorm + AS-norm</b>	<b>ECAPA-TDNN</b>	<b>1.29</b>

NVIDIA A800 GPUs. The learning rate scheduling starts with 10 warm-up epochs with a linear increase from 0 to 0.5, followed by a cosine decay with a final learning rate of 1e-5.

### 4.3. Results and analysis

The experimental results for the VoxCeleb datasets are outlined in Table 1. In the field of self-supervised speaker verification, DINO has been widely adopted, while SDPN represents the latest state-of-the-art model. Both have been reproduced to serve as baseline. Comparing row 1 and 2 shows that SDPN outperforms DINO substantially and consistently across all test sets. The comparison between the results of row 2 and row 3 demonstrates the effectiveness of incorporating off-diagonal dimension regularization in SDPN, achieving EERs of 1.69%, 1.89%, and 3.43%, respectively. Meanwhile, the results from row 2 and row 7 show that adding FroNorm dimension regularization in SDPN further enhances verification performance, achieving relative improvements in EER by 9.44%, 6.53%, and 6.63%, and in MinDCF by 10.8%, 7.63%, and 7.31%, respectively.

The use of score normalization further advances the performance of each system. The addition of Z-norm, T-norm, and AS-norm results in significant performance improvements. Due to the lack of labels in self-supervised systems, score shifts become more pronounced compared to fully supervised speaker verification systems. SDPN with **Frobenius dimension regularization and AS-norm** decreases the EER from 1.80%, 1.99%, and 3.62% to **1.29%**, **1.60%**, and **2.80%**, respectively,

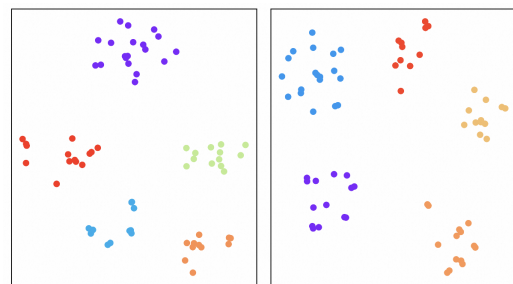


Figure 2: The t-SNE visualization presents extracted embeddings for five speakers, each represented by a different color. The left figure shows embeddings from SDPN, while the right illustrates those from SDPN with dimension regularization. The embeddings with dimension regularization demonstrate improved separation, indicating enhanced discriminability.

with relative improvements of **28.3%**, **19.6%**, and **22.6%** on the three test sets. This represents a substantial performance enhancement, marking a further narrowing of the gap between self-supervised and fully supervised performance.

Additionally, we utilize t-distributed Stochastic Neighbor Embedding (t-SNE) [24] to visually assess the disentanglement performance of speaker embeddings obtained from SDPN and SDPN with Frobenius dimension regularization, as shown in Fig. 2. The embeddings extracted using SDPN with FroNorm dimension regularization demonstrate superior clustering capabilities, indicating that the speaker embeddings are more discriminative. Furthermore, we compare our method with recently proposed self-supervised learning architectures, including those from [4, 6, 9, 10, 12, 16–18], as detailed in Table 2.

## 5. Conclusion

Our work narrows the performance gap between self-supervision and full supervision by introducing self-distillation prototypes network that incorporates dimension regularization and adaptive score normalization. Dimension regularization mitigates the collapse problem by enhancing feature diversity and reducing redundancy. Score normalization effectively tackles the critical issue of score drifting in self-supervised learning. Together, dimension regularization and score normalization enhance the accuracy of speaker verification.

## 6. References

- [1] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, “A simple framework for contrastive learning of visual representations,” in *ICML 2020*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 1597–1607.
- [2] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *CVPR 2020*. Computer Vision Foundation / IEEE, 2020, pp. 9726–9735.
- [3] W. Xia, C. Zhang, C. Weng, M. Yu, and D. Yu, “Self-supervised text-independent speaker verification using prototypical momentum contrastive learning,” in *ICASSP 2021*. IEEE, 2021, pp. 6723–6727.
- [4] X. Zhang, M. Jin, R. Cheng, R. Li, E. Han, and A. Stolcke, “Contrastive-mixup learning for improved speaker verification,” in *ICASSP 2022*. IEEE, 2022, pp. 7652–7656.
- [5] T. Liu, K. A. Lee, Q. Wang, and H. Li, “Disentangling voice and content with self-supervision for speaker recognition,” in *NeurIPS 2023*.
- [6] Y. Tu, M. Mak, and J. Chien, “Contrastive self-supervised speaker embedding with sequential disentanglement,” *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 2704–2715, 2024.
- [7] J. Grill and et al., “Bootstrap your own latent - A new approach to self-supervised learning,” in *NeurIPS 2020*, 2020.
- [8] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging properties in self-supervised vision transformers,” in *ICCV 2021*. IEEE, 2021, pp. 9630–9640.
- [9] 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 *ICASSP 2022*. IEEE, 2022, pp. 6127–6131.
- [10] C. Zhang and D. Yu, “C3-DINO: joint contrastive and non-contrastive self-supervised learning for speaker verification,” *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1273–1283, 2022.
- [11] Y. Chen, S. Zheng, H. Wang, L. Cheng, and Q. Chen, “Pushing the limits of self-supervised speaker verification using regularized distillation framework,” in *ICASSP 2023*. IEEE, 2023, pp. 1–5.
- [12] Y. Chen, S. Zheng, H. Wang, and et al., “Self-distillation prototypes network: Learning robust speaker representations without supervision,” in *ICASSP 2025*.
- [13] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, “Barlow twins: Self-supervised learning via redundancy reduction,” in *ICML 2021*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 12 310–12 320.
- [14] A. Bardes, J. Ponce, and Y. LeCun, “Vicreg: Variance-invariance-covariance regularization for self-supervised learning,” in *ICLR 2022*. OpenReview.net, 2022.
- [15] Q. Garrido, Y. Chen, A. Bardes, L. Najman, and Y. LeCun, “On the duality between contrastive and non-contrastive self-supervised learning,” in *ICLR 2023*.
- [16] H. Heo and et al., “Curriculum learning for self-supervised speaker verification,” in *Interspeech 2023*. ISCA, 2023, pp. 4693–4697.
- [17] Z. Jin, Y. Tu, and M.-W. Mak, “Self-supervised learning with multi-head multi-mode knowledge distillation for speaker verification,” in *Proc. Interspeech*, 2024, pp. 4723–4727.
- [18] Z. Zhao, Z. Li, X. Zhang, W. Wang, and P. Zhang, “Prototype division for self-supervised speaker verification,” *IEEE Signal Process. Lett.*, vol. 31, pp. 880–884, 2024.
- [19] R. Auckenthaler, M. J. Carey, and H. Lloyd-Thomas, “Score normalization for text-independent speaker verification systems,” *Digit. Signal Process.*, vol. 10, no. 1-3, pp. 42–54, 2000.
- [20] H. Aronowitz, D. Irony, and D. Burshtein, “Modeling intra-speaker variability for speaker recognition,” in *INTERSPEECH*. ISCA, 2005, pp. 2177–2180.
- [21] P. M. et al., “Analysis of score normalization in multilingual speaker recognition,” in *Interspeech*. ISCA, 2017, pp. 1567–1571.
- [22] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” in *Interspeech 2018*. ISCA, 2018, pp. 1086–1090.
- [23] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: A large-scale speaker identification dataset,” in *18th Annual Conference of the International Speech Communication Association, Interspeech 2017, Stockholm, Sweden, August 20-24, 2017*. ISCA, 2017, pp. 2616–2620.
- [24] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne.” *Journal of machine learning research*, vol. 9, no. 11, 2008.