Two-stage Semi-supervised Speaker Recognition with Gated Label Learning

Xingmei Wang[1] , **Jiaxiang Meng**[1] , **Kong Aik Lee**[2] , **Boquan Li**[1,3,*] and **Jinghan Liu**[1]
[1]College of Computer Science and Technology, Harbin Engineering University, Harbin, China
[2]Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong
[3]School of Computing and Information Systems, Singapore Management University, Singapore
{wangxingmei, mjxwjy}@hrbeu.edu.cn, kong-aik.lee@polyu.edu.hk, {liboquan, liujinghan}@hrbeu.edu.cn

Abstract

Speaker recognition technologies have been successfully applied in diverse domains, benefiting from the advance of deep learning. Nevertheless, current efforts are still subject to the lack of labeled data. Such issues have been attempted in computer vision, through *semi-supervised learning (SSL)* that assigns pseudo labels for unlabeled data, undertaking the role of labeled ones. Through our empirical evaluations, the state-of-the-art SSL methods show unsatisfactory performance in speaker recognition tasks, due to the imbalance between the quantity and quality of pseudo labels. Therefore, in this work, we propose a two-stage SSL framework, with the aim to address the data scarcity challenge. We first construct an initial contrastive learning network, where the encoder outputs the embedding representation of utterances. Furthermore, we construct an iterative holistic semi-supervised learning network that involves a clustering strategy to assign pseudo labels, and a gated label learning (GLL) strategy to further select reliable pseudo-label data. Systematical evaluations show that our proposed framework achieves superior performance in speaker recognition than the state-of-the-art methods, matching the performance of supervised learning.

1 Introduction

Speaker recognition technologies identify speakers based on their voice features extracted from speaker utterances [Hanifa *et al.*, 2021], which have been broadly utilized in numerous information security applications such as identity authentication and access control [Singh *et al.*, 2012].

With the emergence of deep learning, speaker recognition is implemented as neural networks [Brecht *et al.*, 2020], and the performance is improved into a new level [Bai and Zhang, 2021; Son *et al.*, 2020]. Although remarkable results have been achieved, these methods require massive labeled data to perform *supervised learning* [Vladimir, 2017], and the challenge of annotating utterances as well as huge time expenses make them hard to be applied in practice. In response to such limitations, researchers have attempted strategies including *few-shot learning* [Prateek, 2020; Yanxiong *et al.*, 2023] and *transfer learning* [Cunwei *et al.*, 2018]. Based on their reported results, they still perform limited success if only limited labeled utterances are available.

Intuitively, compared to labeled utterances, the ones without labels are widely available, which inspires a potential direction, i.e., taking advantage of *semi-supervised learning (SSL)* [Engelen and Hoos, 2020] that trains models based on joint labeled as well as unlabeled data. To leverage unlabeled data, typical SSL technologies assign pseudo labels to them enabling such data to act as labeled ones, which has been implemented in computer vision [Xiangli *et al.*, 2023]. Especially, those state-of-the-art holistic methods such as FixMatch [Kihyuk *et al.*, 2020] and FlexMatch [Bowen *et al.*, 2021] achieve promising performance matching supervised learning. In contrast, there are only limited research attempts in speaker recognition [Long *et al.*, 2021; Nakamasa and Keita, 2020; Kreyssig and Woodland, 2020; Fuchuan *et al.*, 2022], and their reported results are inferior to the performance in computer vision. In general, SSL in speaker recognition is still an open problem and has plenty of room for improvement.

In this work, we propose a two-stage holistic SSL framework customized for speaker recognition, and aim to achieve matching performance to supervised learning. To achieve this, the key is to understand *why existing holistic methods perform excellently in computer vision but the same is not true in speaker recognition?* We attribute this question to the fact that classes in utterance data greatly exceed image ones. To be specific, such huge-class data inevitably causes the imbalance between the quality and quantity of pseudo labels, i.e., SSL models focus on either assigning correct labels or selecting enough amounts of pseudo-label data. Such phenomena are analogous to the *confirmation bias* [Eric *et al.*, 2020] issue, i.e., models overfit to the data assigned with incorrect pseudo labels. Based on this intuition, we aim to implement effective speaker recognition by addressing such imbalances.

In particular, our proposed framework mainly includes:

- In Stage I, we devise a contrastive learning-based network [Danwei *et al.*, 2021] (based on unlabeled data), which provides an initial framework as well as produces speaker embedding representations for the next stage.

[*]Corresponding author

- In Stage II, we devise a holistic network to perform iterative semi-supervised learning (based on joint labeled and unlabeled data). First, we apply a semi-supervised clustering strategy to assign pseudo labels, rather than based on classification layers. By performing clustering based on the similarity between labeled and unlabeled embeddings, those incorrect pseudo labels are preliminarily mitigated. Second, we propose a gated label learning (GLL) network that involves flexible threshold and label verification strategies, which balance the quality and quantity of pseudo labels and further select reliable pseudo-label data.
- Based on comparative and ablation experiments, our framework (1) achieves promising speaker recognition performance (1.18%, EER) that approximates supervised learning (0.96%, EER), (2) effectively balances the quality and quantity of pseudo labels, and (3) is superior to the state-of-the-art baseline methods.
- We release our models and codes resulting from this work online[1], and believe this work is pioneer to support future research around SSL-based speaker recognition.

2 Related Work

In this section, we review and compare the related work around speaker recognition and semi-supervised learning.

2.1 Speaker Recognition

Speaker recognition technologies involve identifying and verifying the identity of an individual based on their unique audio characteristics. It is a crucial branch in the domain of artificial intelligence and information security [Singh *et al.*, 2012].

Compared to conventional approaches that perform manual feature engineering [Reynolds *et al.*, 2000; Bai and Zhang, 2021; Noble, 2006; Billson *et al.*, 2019; Campbell *et al.*, 2006], advanced deep learning models adopt Deep Neural Networks (DNNs) to perform speaker recognition that extracts identification features automatically. For example, Snyder et al. [David *et al.*, 2018] proposed a Time Delay Neural Network (TDNN) that contained segment-level as well as time-delay layers, so as to extract time-related features as x-vector for recognition. Desplanques et al. [Brecht *et al.*, 2020] augmented TDNN with a series of Emphasized Channel Attention, Propagation, and Aggregation blocks as ECAPA-TDNN, which learned both temporal and context information for recognition. Deep learning models achieve breakthrough performance benefiting from automatic feature extraction. However, most of them follow supervised learning schemes [David *et al.*, 2021] based on numerous labeled utterances. Expensive annotation expenses make existing work hard to be applied in practice.

In response to the labeled data absence issue, Ali et al. [Yanxiong *et al.*, 2023] devised a few-shot learning method, which designed a feature interaction strategy to enhance the representational ability of its learned embedding. Sun et al. [Cunwei *et al.*, 2018] proposed a Convolutional Neural Network mixed Restricted Boltzmann Machine (TLCNN-RBM) to perform transfer learning. Although these methods mitigate the data annotation issue to certain extents, based on their reported results, limited labeled utterances still make their performance inferior to supervised learning.

[1]Our models, codes and data are available at https://github.com/aitssgll/semi-supervised-speaker-recognition

2.2 Semi-supervised Learning

Semi-supervised learning (SSL) technologies train deep learning models based on few labeled as well as enough unlabeled data. Especially, those typical methods produce pseudo labels for unlabeled data and thereby enable them to undertake the role of labeled ones [Yassine *et al.*, 2020]. We review existing SSL technologies based on their application domains, i.e., computer vision and speaker recognition.

SSL in Computer Vision

Existing methods are generally categorized into three groups, i.e., consistency regularization, entropy minimization, and holistic (the former two types) methods.

First, *consistency regularization methods* [Mehdi *et al.*, 2016] request models' predictions to be consistent across unlabeled as well as augmented data, so as to learn robust and consistent features [Samuli and Timo, 2017; Takeru *et al.*, 2019; Antti and Harri, 2017]. For example, Laine et al. [Samuli and Timo, 2017] proposed a Temporal Ensembling strategy on Pi-Model that adopted an MSE loss to obtain similar predictions between original as well as augmented inputs.

Second, *entropy minimization methods* [Yves and Yoshua, 2005] minimize the entropy of models' predictions, so as to encourage models to produce confident and reliable predictions [Hyun, 2013; Qizhe *et al.*, 2020]. For example, Lee et al. [Hyun, 2013] present pseudo labels that pick up the class with the maximum predicted probability, and can be used as if they are true labels. Xie et al. [Qizhe *et al.*, 2020] proposed a self-training method, which assigned pseudo labels from student-teacher models.

Third, *holistic methods* integrate the above consistent regularization and entropy minimization strategies. Specifically, pseudo labels are assigned under weak augmentation for unlabeled data, and the labeled and pseudo-label data is then jointly utilized with cross-entropy loss [Kihyuk *et al.*, 2020; Bowen *et al.*, 2021]. For example, Sohn et al. [Kihyuk *et al.*, 2020] proposed FixMatch that generated pseudo labels using a model's predictions on weakly-augmented unlabeled data, and the model was further trained under a fixed threshold when fed the strongly-augmented data. Zhang et al. [Bowen *et al.*, 2021] reported that FixMatch utilizes pre-defined constant thresholds for all classes to select unlabeled data, ignoring different learning statuses and difficulties of different classes. They thus proposed FlexMatch that involved a Curriculum Pseudo Labeling (CPL) strategy to obtain the flexible threshold dynamically for each class. Chen et al. [Ting *et al.*, 2020a] proposed SimCLRv2, which adopted contrastive learning with unlabeled data, and was then fine-tuned by a few labeled ones. They found that semi-supervised learning can benefit from two-stage training strategies, especially contrastive learning.

Among the above SSL methods, the holistic ones are acknowledgedly deemed as the state-of-the-art SSL methods, and have reported promising performance close to supervised learning in computer vision tasks [Xiangli *et al.*, 2023].

SSL in Speaker Recognition

Currently, only limited work attempts SSL in the speaker recognition domain. For example, Inoue et al. [Nakamasa and Keita, 2020] proposed a framework based on Generalized Contrastive Loss (GCL), which unified losses from supervised metric learning as well as unsupervised contrastive learning. Kreyssig et al. [Kreyssig and Woodland, 2020] proposed a variant of VAT [Antti and Harri, 2017], where the loss was defined as the robustness of the speaker embedding against input perturbations, and measured by the cosine distance (termed as CD-VAT). Tong et al. [Fuchuan *et al.*, 2022] utilized a Graph Convolutional Network (GCN) to cluster pseudo labels for unlabeled data.

In contrast to the mature SSL technologies in computer vision, existing SS-based speaker recognition methods report inferior results than supervised learning, and there is still plenty of room for improvement. Thus, our work adopts the holistic SSL strategy, given their promising performance in computer vision, and aims to achieve performance as promising as supervised learning.

3 Methodology

In this section, we present our proposed two-stage speaker recognition framework illustrated in Figure 1 in detail. As in the figure, Stage I performs contrastive learning as a pretraining task that provides an initial encoder. Stage II performs iterative holistic semi-supervised learning, where a clustering strategy is first applied on the encoder, so as to assign pseudo labels for unlabeled data based on labeled data. Further, in holistic semi-supervised learning, a cross-entropy loss is jointly utilized for supervised loss L_l (for labeled data x^l) and unsupervised loss L_{u_ft} and L_{u_lv} (for unlabeled data x^u), and the pseudo labels are adopted as supervision signals of the unsupervised loss. Finally, we propose gated labeled learning (GLL) that involves flexible threshold and label verification strategies to further select reliable pseudo-label data. Note that these operations are iteratively conducted.

3.1 Contrastive Learning

Motivated by the conclusion reported in SimCLR-v2 [Ting *et al.*, 2020a], contrastive learning in an upstream task is beneficial to the SSL performance in a downstream task. Such advantages are in accord with our aim to improve the performance of SSL in speaker recognition. Thus, in Stage I, we construct a contrastive learning network that provides an initial SSL framework and produces speaker embeddings.

Specifically, contrastive learning [Danwei *et al.*, 2021; Ting *et al.*, 2020b] trains all unlabeled data in a task-agnostic way with both positive and negative utterance pairs. Formally, the aim is to minimize the distance of the positive pairs:

$$L_{\text{scl}} = \frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{2} - \log \frac{\exp(\cos(e_{i,1}, e_{i,2}))}{\sum\limits_{k=1}^{N} \sum\limits_{l=2}^{2} \mathbb{I}_{\substack{k \neq i \\ j \neq l}} \exp(\cos(e_{i,j}, e_{k,l}))}, \quad (1)$$

where $e_{i,j}$ is the embedding obtained by encoder $f(\cdot)$ on segment $x_{i,j}$, and $cos(\cdot)$ is a specific cosine similarity function.

From Equation 1, to perform contrastive learning, it is necessary to obtain enough positive and negative pairs. However, meaningful information contained in positive pairs is commonly not enough [Ruijie *et al.*, 2022]. Thus, as illustrated in Figure 1, two separate sub-segments $x_{i,1}$ and $x_{i,2}$ are cut randomly from one utterance to positive pairs, so as to enrich the quantity. In contrast, two segments from different utterances are viewed as negative pairs. Moreover, to avoid false-negative pairs, we adopt large-enough datasets and suitable batch sizes, which are validated to reduce false-negative rates [Haoran *et al.*, 2021]. Additionally, our strategy of contrastive learning follows [Ting *et al.*, 2020b], where an Augmentation Adversarial Training (AAT) loss [Jaesung *et al.*, 2020] is jointly utilized with the contrastive loss, so as to maintain the classification ability of encoder $f(\cdot)$, whilst reducing its mislabeling.

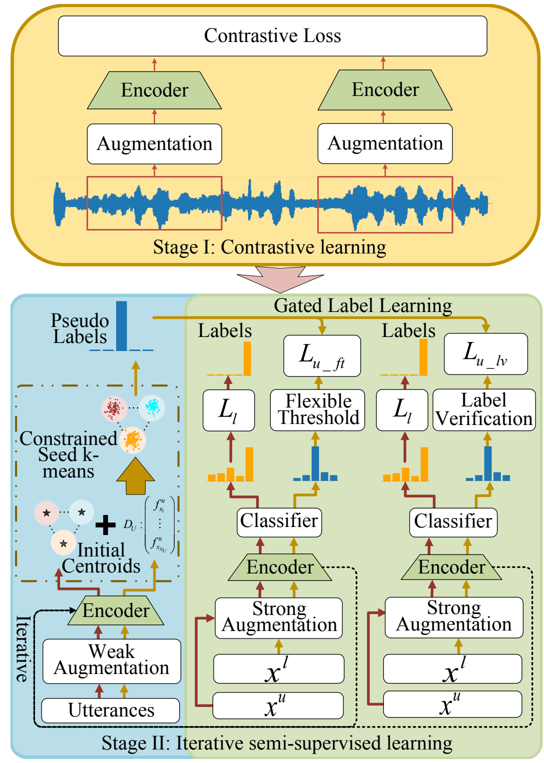


Figure 1: The proposed two-stage semi-supervised speaker recognition framework, where Stage I (top) performs contrastive learning and Stage II (bottom) performs iterative holistic semi-supervised learning with gated label learning.

3.2 Iterative Holistic Semi-supervised Learning

Based on the initial network in Stage I, we construct a further SSL framework for speaker recognition in Stage II. Following our analysis in Section 2.2, we adopt the holistic SSL strategy. It generally utilizes labeled data with ground-truth labels as well as abundant unlabeled data with pseudo labels produced under weak augmentation, and the pseudo label is a breakthrough factor to ensure reliable performance [Eric *et al.*, 2020]. Thus, we propose two strategies to assign and select reliable pseudo labels, i.e., clustering and gated label learning.

Semi-supervised Clustering

Based on labeled and unlabeled utterances, a constrained seed k-means [Sugato, 2002] clustering strategy is adopted to assign pseudo labels for the unlabeled utterance, upon the similarity between unlabeled embedding e_i^u and labeled embedding e_j^l. This design preliminarily avoids the mislabeling of pseudo labels for unlabeled utterances [Hieu *et al.*, 2021], instead of assigning them based on classification layers in conventional methods [Kihyuk *et al.*, 2020; Bowen *et al.*, 2021].

The framework is then trained based on labeled utterances with ground-truth labels as well as unlabeled ones with pseudo labels. Moreover, an additive angular margin softmax (AAM-softmax) loss [Jiankang *et al.*, 2022] is augmented to the encoder $f(\cdot)$. Note that we adopt an iterative learning strategy. That is, the speaker encoder is iteratively trained with multiple iterations until converges, and the best-performed parameters in certain iterations are then utilized for clustering, which reassigns pseudo labels in the next iteration.

Gated Label Learning

Next, in Stage II, we propose an additional gated label learning (GLL) strategy that involves flexible threshold and label verification strategies, to further select reliable pseudo-label data and balance the quality and quantity of pseudo labels.

General methods select pseudo labels based on either fixed [Kihyuk *et al.*, 2020] or flexible thresholds [Bowen *et al.*, 2021; Yidong *et al.*, 2022]. Although such methods achieve satisfactory performance in computer vision tasks, if high-quality (correct) pseudo labels are over-focused, it inevitably reduces the quantity of selected labels and affects the performance of SSL. In speaker recognition tasks, the quantity and quality of pseudo labels are more challenging to be balanced compared with image ones [Hao *et al.*, 2023; Nayeem *et al.*, 2021]. Thus, we propose GLL that provides a fusion way to filter pseudo labels.

Preliminarily, the *quality* refers to the ratio of the correctly assigned pseudo labels compared with their ground-truth ones:

$$quality = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(y_i^{pc} == y_u), \tag{2}$$

where y_i^{pc} is the pseudo labels of the unlabeled samples through GLL, y_u represents their ground truth labels, and N is the number of unlabeled samples selected by GLL.

The *quantity* refers to the ratio of the selected pseudo-label data among the total unlabeled ones:

$$quantity = \frac{1}{N_U}\sum_{i=1}^{N_U}\mathbb{I}\left(B\left(i\right)\right), \tag{3}$$

$$B\left(\cdot\right) = \begin{cases} q_i^u > \tau & GLL == flexiblethreshold \\ \arg\max\left(q_i^u\right) = y_i^{pc} & otherwise \end{cases}, \tag{4}$$

where y_i^{pc} is the pseudo labels assigned by clustering, N_u is the number of the total unlabeled samples, and τ is the confidence threshold. q_i^u is the speaker recognition prediction on data x_i^u, that is, $q_i^u = p\left(y|A_W\left(x_i^u\right)\right)$. $p\left(\cdot|\cdot\right)$ is the output probability, where $A_W\left(\cdot\right)$ represents a weak augmentation operation. Note that the ground truth labels of the unlabeled sample in Equation 2 are only used to perform analysis, instead of guiding model training.

First, we apply flexible thresholds [Bowen *et al.*, 2021] into the confidence of the loss function, so as to apply different thresholds to select pseudo-label data. Formally, the flexible threshold is:

$$\tau_t = \begin{cases} \frac{1}{K}, & t = 0 \\ \mu\tau_{t-1} + (1-\mu)\frac{1}{\mu B}\sum\limits_{j=1}^{\mu B}\mathbb{I}(q_i^u > \tau_{t-1})\cdot q_i^u, & otherwise \end{cases}, \tag{5}$$

where μB is the batch size and μ is the hyperparameter to balance the increasing speed of threshold τ_t. Here, the unsupervised loss of the flexible threshold is:

$$L_{u\_ft} = \frac{1}{N_U}\sum_{i=1}^{N_U}\left(\mathbb{I}(q_i^u > \tau_t)\cdot H\left(y_i^{pc}, p\left(y|A_S\left(x_i^u\right)\right)\right)\right), \tag{6}$$

where $H\left(\cdot,\cdot\right)$ is the cross-entropy loss, and $A_S\left(\cdot\right)$ represents a strong augmentation operation.

Second, we apply label verification as a decision fusion to the loss function, so as to enable pseudo labels to be assigned by clustering or classification. Such strategies provide flexible mechanisms to avoid incorrect pseudo labels caused by poor classification ability at the beginning of model training. Formally, the unsupervised loss of label verification is:

$$L_{u\_lv} = \frac{1}{N_U}\sum_{i=1}^{N_U}\mathbb{I}\left(\arg\max\left(q_i^u\right) = y_i^{pc}\right)\cdot H\left(y_i^{pc}, p\left(y|A_S\left(x_i^u\right)\right)\right). \tag{7}$$

Finally, the overall loss as:

$$L = L_l + \lambda L_u, \tag{8}$$

where L_l is a cross-entropy loss (for training labeled data) and L_u is the unsupervised training loss involving L_{u_ft} and L_{u_lv}. Note that they are chosen alternatively to select reliable pseudo labels until the next re-clustering.

4 Experiment

In this section, we first introduce our experimental setup, and then present our comparative as well as ablation experiments.

4.1 Experimental Setup

We start with introducing the datasets, implementation, and baseline methods of our experiments.

Datasets

To train our framework, we adopt the most typical datasets, VoxCeleb2 [Son *et al.*, 2018], which contains 1092009 utterances from 5994 speakers. In addition, we collect the testing set from VoxCeleb1 [Arsha *et al.*, 2017], which contains 37721 utterance pairs from 40 speakers. Following the setting of typical SSL speaker recognition methods [Long *et al.*, 2021] and the general settings of SSL, i.e., the quantity of unlabeled data should be more than labeled ones, different proportions of utterances per speaker (1% (1 sample), 2% (4 samples), 6% (10 samples), 11% (20 samples), 22% (40 samples), 33% (60 samples)) are selected as labeled data, and the remaining utterances are selected as unlabeled ones.

Implementation

In Stage I, we construct our contrastive learning network based on a Loss-gated Learning (LGL) [Ting *et al.*, 2020b] architecture (one of the state-of-the-art self-supervised models), following the parameter settings in ECAPA-TDNN [Brecht *et al.*, 2020], which is one of the state-of-the-art end-to-end

Method	The proportion of utilized labeled data					
	1%	2%	6%	11%	22%	33%
FlexMatch (0.9) [Bowen *et al.*, 2021]	21.22	10.05	6.68	8.21	6.83	6.43
FlexMatch (0.000351) [Bowen *et al.*, 2021]	19.89	16.53	18.57	6.99	7.49	1.24
FixMatch (0.9) [Kihyuk *et al.*, 2020]	15.62	10.24	9.07	8.44	8.04	7.74
FixMatch (0.000351) [Kihyuk *et al.*, 2020]	13.79	9.33	5.27	2.82	2.72	2.53
Mean Teacher [Takeru *et al.*, 2019]	32.62	10.36	6.35	3.74	1.95	1.79
Pseudo Label [Hyun, 2013]	17.96	11.51	7.11	3.28	6.91	6.52
SimCLRv2 [Ting *et al.*, 2020a]	6.05	3.16	2.97	2.66	2.59	2.03
GCL [Nakamasa and Keita, 2020]	2.56					
CD-VAT [Kreyssig and Woodland, 2020]	6.46					
GCN [Fuchuan *et al.*, 2022]	1.30					
Ours Stage I only	6.61					
Ours w/o GLL	4.58	2.61	2.15	1.81	1.64	1.37
Ours	**3.24**	**1.74**	**1.65**	**1.53**	**1.41**	**1.18**
Full supervised	0.96					

Table 1: EER (%) results of our framework as well as other baseline methods

networks in speaker recognition [Chen *et al.*, 2023]. Specifically, the channel size of ECAPA-TDNN is 1024, and the log mel-spectrogram dimension is 80. In Stage II, we apply strong augmentation settings in X-Vectors [David *et al.*, 2018], and set weak augmentation as no data is augmented. The clustering component is implemented based on a *faiss library* [Mathilde *et al.*, 2018]. Finally, the optimizer in both stages is Adam [Kingma and Lei, 2015] with an initial learning rate of 0.001. The learning rate is decreased by 5% each five epochs in Stage I and is decreased by 5% each epoch in Stage II.

Baseline Methods

As illustrated in the first (Method) column in Table 1, to evaluate the effectiveness of our framework, we apply multiple state-of-the-art SSL methods in both computer vision and speaker recognition domains. Computer vision methods include the holistic FixMatch [Kihyuk *et al.*, 2020] and FlexMatch [Bowen *et al.*, 2021], the consistency-regularization Mean Teacher [Takeru *et al.*, 2019], the entropy-minimization Pseudo Label [Hyun, 2013], and SimCLRv2 [Ting *et al.*, 2020a] that inspires us from applying the contrastive learning in Stage I. Speaker recognition methods include GCL [Nakamasa and Keita, 2020], CD-VAT [Kreyssig and Woodland, 2020] and GCN [Fuchuan *et al.*, 2022].

Note that the datasets utilized in speaker recognition and computer vision tasks are greatly different, especially their contained classes. For example, 5994 classes are contained in VoxCeleb2 [Son *et al.*, 2018] and about 1000 classes are contained in ImageNet [Jia *et al.*, 2009]. Thus, for fair comparisons, we explore the best settings for those computer vision methods (to adapt to speaker recognition datasets), by adjusting the confidence threshold.

Specifically, we explore the confidence threshold based on the state-of-the-art holistic method, FixMatch, which commonly sets the threshold to 0.9, and the setting is validated effectively in computer vision tasks. By evaluating FixMatch in our speaker recognition dataset (based on 6% labeled data), we observe that the maximum confidence will not exceed 0.0005 and the maximized mean confidence is 0.00039. We apply the obtained maximized mean confidence (0.00039) as a standard, and multiply it by 0.9 to obtain the best threshold, 0.000351. We also attempt multiple thresholds and find that 0.000351 achieves the best performance in terms of EER (6.49%), and the value is thus selected in the subsequent experiments.

4.2 Comparative Experiments

In the following, we first evaluate and compare our framework with other baselines. We then analyze the quality and quantity of their pseudo labels to empirically explore the reason for their success or failure.

Speaker Recognition Performance

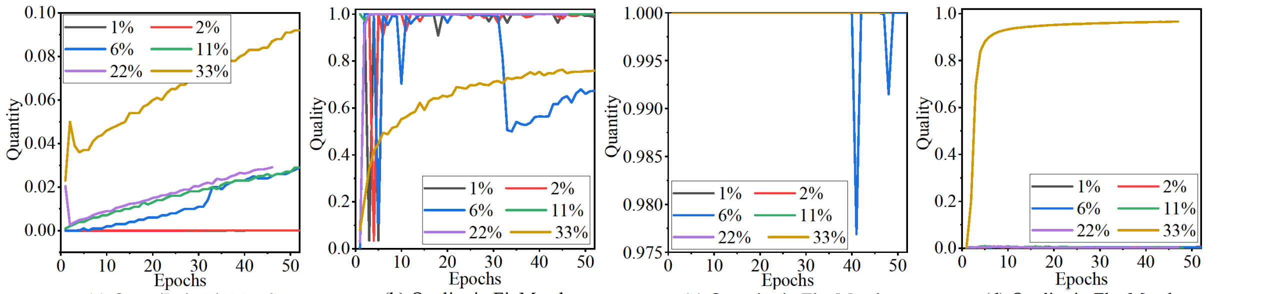
Table 1 presents our results based on the metric of Equal Error Rate (EER), which is commonly adopted to evaluate speaker recognition models [Karen and Andrew, 2015; Chen *et al.*, 2023]. In the table, the results of SSL methods are obtained with different proportions of utilized labeled data, and the results of GCL, CD-VAT, and GCN are referred from their original literature.

First, it is observed that our frameworks achieve the best results (as the blue-highlighted values in the table) that outperform any baseline methods as well as our variants. Especially, based on 1% labeled data, we achieve an EER of 3.24%, which is worthy emphasized that our framework achieves promising results with only few labeled data. Moreover, when 33% labeled data is utilized, our framework achieves an approximate EER (1.18%) to the full supervised model (0.96%), which presents great effectiveness and advancements.

Second, SimCLRv2 outperforms others (except our framework), suggesting that the two-stage framework that inspires us is effective in speaker recognition. FixMatch and FlexMatch perform better based on the confidence threshold of 0.000351 than 0.9 in most cases. Moreover, FixMatch and FlexMatch achieve generally better performance than Mean Teacher and Pseudo Label, which demonstrates the superiority of such holistic methods.

Third, Ours w/o GLL achieves better EER even if only 1% labeled data is utilized (4.58%), which indicates our clustering strategy make contributions to the performance.

In general, our framework is effective in the speaker recognition task even if only limited labeled data is available, and is



(a) Quantity in FixMatch (b) Quality in FixMatch (c) Quantity in FlexMatch (d) Quality in FlexMatch

Figure 2: (a) Quantity and (b) Quality of selected pseudo labels in FixMatch. (c) Quantity and (d) Quality of selected pseudo labels in FlexMatch.

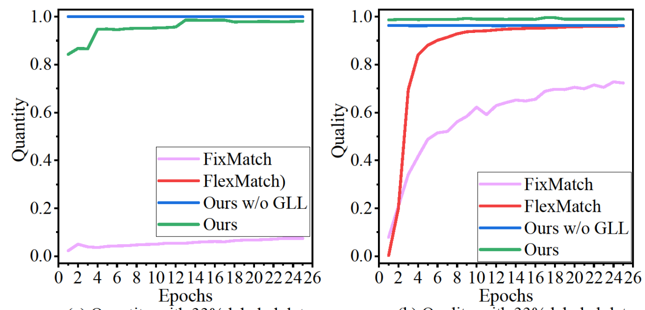
superior to the state-of-the-art methods.

Quality and Quantity of Pseudo Labels

The above experiments demonstrate our framework is superior to those baseline methods. Recall that we attribute our success to the balance of the quality and quantity of pseudo labels. Thus, in this experiment, we first take the holistic methods, FixMatch (0.000351) and FlexMatch (0.000351), as baselines to evaluate whether their pseudo labels are imbalanced, and then compare them with our framework.

Figure 2 presents the quality and quantity results on FixMatch and FlexMatch, where *quality* and *quantity* are defined in Equation 2 and Equation 3. In these figures, the quantity and quality changes are evaluated based on utilizing different amounts of labeled data, along with training epochs. First, to observe the results of FixMatch in Figure 2 (b), high-quality pseudo labels are obtained in most cases. However, in Figure 2 (a), the quantity of the selected pseudo-label data is below 0.1 in all cases. Such results confirm our intuition that FixMatch is over-focusing on assigning correct pseudo labels, which results in low quantity, i.e., the selected pseudo-label data for model training is insufficient. Second, to observe the result of FlexMatch in Figure 2 (c), the pseudo labels are high-quantity in all cases, which indicates the utilization rate of unlabeled data is satisfactory. However, in Figure 2 (d), the quality of pseudo labels is unsatisfactory in most cases, which indicates most of the labels are mislabeled. As an exception, as the yellow curve in Figure 2 (d), high-quality pseudo labels can be assigned with training iterations. This is explainable since 33% labeled data is enough for FlexMatch to assign pseudo labels based on its classification layers. In general, the evaluation results are in accordance with our intuitions, that is, such state-of-the-art SSL models can not balance the quality and quantity of pseudo labels when encountering speaker datasets, and are thus not qualified for the speaker recognition task.

Figure 3 presents the comparison results of FixMatch, FlexMatch, and our frameworks. Compared with FixMatch and FlexMatch, our frameworks (with and without GLL) have achieved both high-quality and high-quantity results. Especially, as the green curve in Figure 3 (b), the quality results achieved by our framework (with GLL) are approaching 1.0, which indicates most of the pseudo labels are correctly labeled. Moreover, as the green curve in Figure 3 (a), the success quality results (in Figure 3 (b)) are obtained based on acceptable quantity drops, which are approaching 1.0 after the 14-th epoch. In general, the results have explained the success of our framework, which has the ability to balance the two factors, and is competent for the speaker recognition task.



(a) Quantity with 33% labeled data (b) Quality with 33% labeled data

Figure 3: Comparisons of the (a) Quantity and (b) Quality of selected pseudo labels in baseline and our frameworks. Note that the red and blue curves in (a) are coincident and are thus occluded by each other.

4.3 Ablation Experiments

Next, we perform ablation experiments to explore the effect of our designed components or strategies on the performance of speaker recognition. Specifically, we first evaluate the flexible threshold and label verification strategies in GLL, and then analyze the impact of iterative learning.

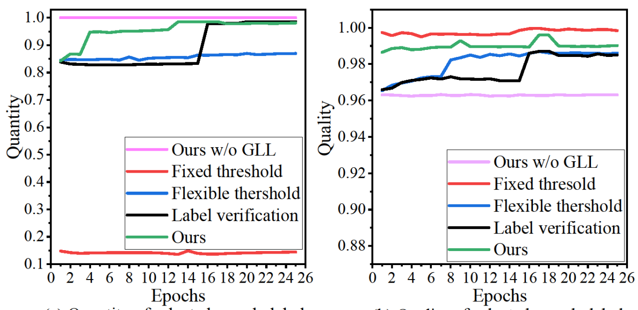
Flexible Threshold and Label Verification

We have implemented four variants of our framework as baselines, their performance is still evaluated on EER, and the quantity and quality of pseudo labels.

Table 2 presents EER results based on different proportions of labeled data. It is first observed that our framework (with GLL) performs best among these baselines in all cases, as the blue-highlighted values in the tables. Second, our framework without GLL generally performs the worst among these methods, which indicates the procedure of selecting pseudo-label data is necessary for the task. For example, based on 1% labeled data, it achieves an EER of 4.58% that is worse than the results of other methods (3.67%, 3.30%, 3.43%, and 3.24%). Third, our framework with either flexible thresholds or label verification, outperforms the fixed-threshold one. For example, based on 2% labeled data, our framework with a fixed thresh-

Method	The proportion of utilized labeled data					
	1%	2%	6%	11%	22%	33%
Ours w/o GLL	4.58	2.61	2.15	1.81	1.64	1.37
Fixed threshold	3.67	2.24	2.13	1.79	1.58	1.54
Flexible threshold	3.30	2.57	2.02	1.75	1.49	1.50
Label verification	3.43	2.36	1.91	1.69	1.53	1.20
Ours	**3.24**	**1.74**	**1.65**	**1.53**	**1.41**	**1.18**

Table 2: EER (%) results of our variant methods



(a) Quantity of selected pseudo labels (b) Quality of selected pseudo labels

Figure 4: (a) Quantity and (b) Quality of selected pseudo labels in variant methods.

old achieves an EER of 2.24%, which is higher than the rest two methods (2.57% and 2.36%). Such results are in accord with our intuition that the fixed-threshold strategy is challenging to assign pseudo labels thus affecting speaker recognition performance, and we will further analyze it in the following experiments. Fourth, the superiority of flexible thresholds and label verification is optimal as they are jointly utilized in our framework (with GLL), which demonstrates the rationality of our designed GLL. In general, the evaluation results have further confirmed the effectiveness of our proposed GLL, and each of its involved strategies is indispensable.

Figure 4 presents the quantity and quality of pseudo labels, and the results are obtained using 33% labeled data. Firstly, as the blue curve in Figure 4 (b), the framework with a fixed threshold achieves high-quality results, however, in Figure 4 (a), it performs extremely low quantity results. Such results have explained their unremarkable EER performance in Table 2, i.e., it over-focuses on the correctness of pseudo labels resulting in the selected pseudo-label data being insufficient. In contrast, as the green curves in Figure 4, our framework achieves both satisfactory quality and quantity results, especially after the 14-th epoch, indicating it has balanced the two factors and has selected reliable pseudo to perform SSL in speaker recognition. In general, these curves are in accordance with the EER results in Table 2, which provides further evidence to prove the necessity of our designed strategies.

Iterative Learning

Recall that our proposed iterative learning strategy iteratively performs clustering and GLL progresses for selecting optimal pseudo labels. In Table 3, we compare the performance of our framework as well as its variants in five iterations, i.e., evaluating their effectiveness with iterative learning.

It is first observed that each result in the second to the fifth

Labels	Method	Iteration				
		1	2	3	4	5
1%	Ours w/o GLL	5.13	**4.58**	4.71	7.62	6.78
	Ours	4.40	3.78	3.48	3.31	**3.24**
2%	Ours w/o GLL	3.31	2.80	**2.61**	2.70	2.75
	Ours	2.80	1.96	**1.74**	1.91	2.03
6%	Ours w/o GLL	3.00	2.51	**2.15**	2.26	2.23
	Ours	2.17	1.88	1.74	**1.65**	1.92
11%	Ours w/o GLL	2.36	1.83	**1.81**	1.91	2.14
	Ours	1.85	**1.53**	1.69	1.82	1.90
22%	Ours w/o GLL	2.23	1.73	1.67	1.68	**1.64**
	Ours	1.69	1.46	**1.41**	1.64	1.70
33%	Ours w/o GLL	1.96	1.52	1.46	**1.37**	1.59
	Ours	1.39	1.28	1.22	**1.18**	1.32

Table 3: EER (%) results of our frameworks in different iterations

iteration outperforms the first iteration, whether our framework is with or without GLL. For example, based on 1% labeled data, our framework (with GLL) achieves an EER of 4.40% in the first iteration, and the results are respectively 3.78%, 3.48%, 3.31% and 3.24% in the rest ones, which indicates our strategy of iteratively assigning and selecting pseudo labels are beneficial to the task. Moreover, we find that the best results (as the blue-highlighted values in the table) commonly appear in the middle (second to fourth) iterations. Based on 2% labeled data, the best EER of our framework (with GLL) appears in the third iteration (1.74%), which indicates the best (quality and quantity) pseudo labels can be selected within limited iterations. However, the degradation in Table 3 as iterations go on is a normal phenomenon in deep learning models, since the model inevitably overfits to labeled data. In general, the experimental results have demonstrated our proposed iterative learning strategies are effective.

5 Conclusion and Future Work

This work proposes a two-stage semi-supervised speaker recognition framework, towards overcoming the challenge posed by limited labeled data, and achieving promising performance matching supervised learning. Specifically, we (1) construct an initial network trained on contrastive learning, (2) apply clustering strategies to produce pseudo labels for unlabeled data, and (3) propose a gated label learning (GLL) network to select reliable pseudo-label data. Experimental results show that our framework is (1) superior to the state-of-the-art methods, (2) explainable by balancing the quantity and quality of pseudo labels, and (3) comparable to the outstanding performance of SSL in computer vision. In the future, achieving promising results with the least labeled data and fine-tuning an broader SSL method to the audio-related tasks are still interesting topics to be perseverely explored.

Acknowledgments

This work was supported by Fundamental Research Funds for the Central Universities in China (3072022JC0601, 3072024CFJ0601), the Primary Research & Developement Plan of Heilongjiang Province (No.GA23A903), and the China Scholarship Council program (Project ID:202306680025)

References

[Antti and Harri, 2017] Tarvainen Antti and Valpola Harri. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, pages 1196–1205, Long Beach, CA, United states, 2017.

[Arsha *et al.*, 2017] Nagraniy Arsha, Chungy Joon Son, and Zisserman Andrew. Voxceleb: A large-scale speaker identification dataset. In *INTERSPEECH*, pages 2616–2620, Stockholm, Sweden, 2017.

[Bai and Zhang, 2021] Zhongxin Bai and Xiao-Lei Zhang. Speaker recognition based on deep learning: An overview. *Neural Networks*, 140:65–99, 2021.

[Billson *et al.*, 2019] Mokgonyane T. Billson, Sefara T. Joseph, Modipa T. Isaiah, and Manamela M. Jonas. Automatic speaker recognition system based on optimised machine learning algorithms. In *IEEE AFRICON Conference*, Accra, Ghana, 2019.

[Bowen *et al.*, 2021] Zhang Bowen, Wang Yidong, Hou Wenxin, Wu Hao, Wang Jindong, Okumura Manabu, and Shinozaki Takahiro. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *NeurIPS*, volume 22, pages 18408–18419, Virtual, Online, 2021.

[Brecht *et al.*, 2020] Desplanques Brecht, Thienpondt Jenthe, and Demuynck Kris. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *INTERSPEECH*, pages 3830–3834, Shanghai, China, 2020.

[Campbell *et al.*, 2006] William Campbell, Douglas E. Sturim, and Douglas A Reynolds. Support vector machines using gmm supervectors for speaker verification. *IEEE Signal Processing Letters*, 13(5):308–311, 2006.

[Chen *et al.*, 2023] Zhengyang Chen, Han Bing, Xiang Xu, Huang Houjun, Liu Bei, and Qian Yanmin. Build a sre challenge system: Lessons from voxsrc 2022 and cnsrc 2022. In *INTERSPEECH*, pages 3202–3206, Dublin, Ireland, 2023.

[Cunwei *et al.*, 2018] Sun Cunwei, Yang Yuxin, Wen Chang, Xie Kai, and Wen Fangqing. Voiceprint identification for limited dataset using the deep migration hybrid model based on transfer learning. *Sensors*, 18(7):2399, 2018.

[Danwei *et al.*, 2021] Cai Danwei, Wang Weiqing, and Li Ming. An iterative framework for self-supervised deep speaker representation learning. In *ICASSP*, pages 6728–6732, Virtual, Toronto, ON, Canada, 2021.

[David *et al.*, 2018] Snyder David, Garcia-Romero Daniel, Sell Gregory, Povey Daniel, and Khudanpur Sanjeev. X-vectors: Robust dnn embeddings for speaker recognition. In *ICASSP*, pages 5329–5333, Calgary, AB, Canada, 2018.

[David *et al.*, 2021] Sztaho David, Szaszak Gyorgy, and Beke Andras. Deep learning methods in speaker recognition: A review. *Periodica polytechnica Electrical engineering and computer science*, 65(4):310–328, 2021.

[Engelen and Hoos, 2020] Jesper Van Engelen and Holger Hoos. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440, 2020.

[Eric *et al.*, 2020] Arazo Eric, Ortego DIego, Albert Paul, O'Connor Noel E., and McGuinness Kevin. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *IJCNN*, pages 1–8, Virtual, Glasgow, United kingdom, 2020.

[Fuchuan *et al.*, 2022] Tong Fuchuan, Zheng Siqi, Zhang Min, Chen Yafeng, Suo Hongbin, Hong Qingyang, and Li Lin. Graph convolutional network based semi-supervised learning on multi-speaker meeting data. In *ICASSP*, pages 6622–6626, Virtual, Online, Singapore, 2022.

[Hanifa *et al.*, 2021] Rafizah Mohd Hanifa, Khalid Isa, and Shamsul Mohamad. A review on speaker recognition: Technology and challenges. *Computers & Electrical Engineering*, 90:107005, 2021.

[Hao *et al.*, 2023] Chen Hao, Tao Ran, Fan Yue, Wang Yidong, Wang Jindong, Schiele Bernt, Xie Xing, Raj Bhiksha, and Savvides Marios. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. *arXiv preprint arXiv:2301.10921*, 2023.

[Haoran *et al.*, 2021] Zhang Haoran, Zou Yuexian, and Wang Helin. Contrastive self-supervised learning for text-independent speaker verification. In *ICASSP*, pages 6713–6717, Virtual, Toronto, ON, Canada, 2021.

[Hieu *et al.*, 2021] Pham Hieu, Dai Zihang, Xie Qizhe, and Le Quoc V. Meta pseudo labels. In *CVPR*, pages 11553–11563, Virtual, Online, United states, 2021.

[Hyun, 2013] Lee Dong Hyun. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML*, volume 3, page 896, 2013.

[Jaesung *et al.*, 2020] Huh Jaesung, Heo Hee Soo, Kang Jingu, Watanabe Shinji, and Chung Joon Son. Augmentation adversarial training for unsupervised speaker recognition. *arXiv preprint arXiv:2007.12085*, 2020.

[Jia *et al.*, 2009] Deng Jia, Dong Wei, Socher Richard, Li Li-Jia, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, Miami, FL, USA, 2009.

[Jiankang *et al.*, 2022] Deng Jiankang, Guo Jia, Yang Jing, Xue Niannan, Kotsia Irene, and Zafeiriou Stefanos. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, 2022.

[Karen and Andrew, 2015] Simonyan Karen and Zisserman Andrew. Very deep convolutional networks for large-scale image recognition. In *ICLR*, San Diego, CA, United states, 2015.

[Kihyuk *et al.*, 2020] Sohn Kihyuk, Berthelot David, Li Chun-Liang, Zhang Zizhao, Carlini Nicholas, Cubuk Ekin D., Kurakin Alex, Zhang Han, and Raffel Colin. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, pages 596–608, Virtual, Online, 2020.

[Kingma and Lei, 2015] Diederik Pieter Kingma and Ba Jimmy Lei. Adam: A method for stochastic optimization. In *ICLR*, San Diego, CA, United states, 2015.

[Kreyssig and Woodland, 2020] Florian Kreyssig and Philip Woodland. Cosine-distance virtual adversarial training for semi-supervised speaker-discriminative acoustic embeddings. In *INTERSPEECH*, pages 3241–3245, Shanghai, China, 2020.

[Long *et al.*, 2021] Chen Long, Ravichandran Venkatesh, and Stolcke Andreas. Graph-based label propagation for semi-supervised speaker identification. In *INTERSPEECH*, volume 4, pages 2583–2587, Brno, Czech republic, 2021.

[Mathilde *et al.*, 2018] Caron Mathilde, Bojanowski Piotr, Joulin Armand, and Douze Matthijs. Deep clustering for unsupervised learning of visual features. In *ECCV*, pages 139–156, Munich, Germany, 2018.

[Mehdi *et al.*, 2016] Sajjadi Mehdi, Javanmardi Mehran, and Tasdizen Tolga. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *NeurIPS*, volume 0, pages 1171–1179, Barcelona, Spain, 2016.

[Nakamasa and Keita, 2020] Inoue Nakamasa and Goto Keita. Semi-supervised contrastive learning with generalized contrastive loss and its application to speaker recognition. In *APSIPA ASC*, pages 1641–1646, Auckland, New zealand, 2020.

[Nayeem *et al.*, 2021] Rizve Mamshad Nayeem, Duarte Kevin, Rawat Yogesh S, and Shah Mubarak. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *ICLR*, Virtual, Online, 2021.

[Noble, 2006] William Stafford Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567, 2006.

[Prateek, 2020] Mishra Prateek. Few shot text-independent speaker verification using 3d-cnn. *arXiv preprint arXiv:2008.11088*, 2020.

[Qizhe *et al.*, 2020] Xie Qizhe, Luong Minh-Thang, Hovy Eduard, and Le Quoc V. Self-training with noisy student improves imagenet classification. In *CVPR*, pages 10684–10695, Virtual, Online, United states, 2020.

[Reynolds *et al.*, 2000] Douglas A Reynolds, Thomas F Quatieri, and Robert B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1):19–41, 2000.

[Ruijie *et al.*, 2022] Tao Ruijie, Lee Kong Aik, Das Rohan Kumar, Hautamäki Ville, and Li Haizhou. Self-supervised speaker recognition with loss-gated learning. In *ICASSP*, pages 6142–6146, 2022.

[Samuli and Timo, 2017] Laine Samuli and Aila Timo. Temporal ensembling for semi-supervised learning. In *ICLR*, Toulon, France, 2017.

[Singh *et al.*, 2012] Nilu Singh, R.A. Khan, and Raj Shree. Applications of speaker recognition. *Procedia Engineering*, 38:3122–3126, 2012.

[Son *et al.*, 2018] Chung Joon Son, Nagrani Arsha, and Zisserman Andrew. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, pages 1086–1090, Hyderabad, India, 2018.

[Son *et al.*, 2020] Chung Joon Son, Huh Jaesung, Mun Seongkyu, Lee Minjae, Heo Hee Soo, Choe Soyeon, Ham Chiheon, Jung Sunghwan, Lee Bong-Jin, and Han Icksang. In defence of metric learning for speaker recognition. In *INTERSPEECH*, pages 2977–2981, Shanghai, China, 2020.

[Sugato, 2002] Basu Sugato. Semi-supervised clustering by seeding. In *ICML*, 2002.

[Takeru *et al.*, 2019] Miyato Takeru, Maeda Shin-Ichi, Koyama Masanori, and Ishii Shin. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2019.

[Ting *et al.*, 2020a] Chen Ting, Kornblith Simon, Swersky Kevin, Mohammad Norouzi, and Hinton Geoffrey. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, pages 22243–22255, Virtual, Online, 2020.

[Ting *et al.*, 2020b] Chen Ting, Kornblith Simon, Norouzi Mohammad, and Hinton Geoffrey. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1575–1585, Virtual, Online, 2020.

[Vladimir, 2017] Nasteski Vladimir. An overview of the supervised machine learning methods. *Horizons. b*, 4:51–62, 2017.

[Xiangli *et al.*, 2023] Yang Xiangli, Song Zixing, King Irwin, and Xu Zenglin. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8934–8954, 2023.

[Yanxiong *et al.*, 2023] Li Yanxiong, Chen Hao, Cao Wenchang, Huang Qisheng, and He Qianhua. Few-shot speaker identification using lightweight prototypical network with feature grouping and interaction. *IEEE Transactions on Multimedia*, pages 1–12, 2023.

[Yassine *et al.*, 2020] Ouali Yassine, Hudelot Céline, and Tami Myriam. An overview of deep semi-supervised learning. *arXiv preprint arXiv:2006.05278*, 2020.

[Yidong *et al.*, 2022] Wang Yidong, Chen Hao, Heng Qiang, Hou Wenxin, Fan Yue, Wu Zhen, Wang Jindong, Savvides Marios, Shinozaki Takahiro, Raj Bhiksha, Schiele Bernt, and Xie Xing. Freematch: self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022.

[Yves and Yoshua, 2005] Grandvalet Yves and Bengio Yoshua. Semi-supervised learning by entropy minimization. In *NeurIPS*, pages 529–536, Vancouver, BC, Canada, 2005.