Towards Regularized Mixture of Predictions for Class-Imbalanced Semi-Supervised Facial Expression Recognition

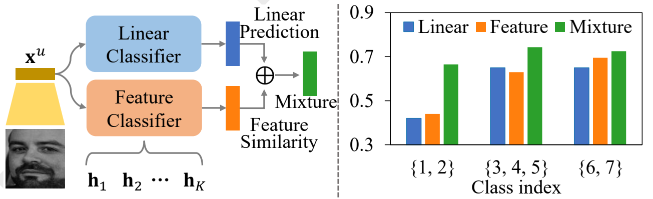
Hangyu Li[1] , **Yixin Zhang**[2] , **Jiangchao Yao**[3] , **Nannan Wang**[2,*] and **Bo Han**[1]
[1]TMLR Group, Department of Computer Science, Hong Kong Baptist University, Hong Kong, China
[2]State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, China
[3]Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai, China
{hangyuli.xidian, yxzhang.xidian}@gmail.com, Sunarker@sjtu.edu.cn, nnwang@xidian.edu.cn,
bhanml@comp.hkbu.edu.hk

Abstract

Semi-supervised facial expression recognition (SSFER) effectively assigns pseudo-labels to confident unlabeled samples when only limited emotional annotations are available. Existing SSFER methods are typically built upon an assumption of the class-balanced distribution. However, they are far from real-world applications due to biased pseudo-labels caused by class imbalance. To alleviate this issue, we propose **Re**gularized **M**ixture **o**f **P**redictions (ReMoP), a simple yet effective method to generate high-quality pseudo-labels for imbalanced samples. Specifically, we first integrate feature similarity into the linear prediction to learn a mixture of predictions. Furthermore, we introduce a class regularization term that constrains the feature geometry to mitigate imbalance bias. Being practically simple, our method can be integrated with existing semi-supervised learning and SSFER methods to tackle the challenge associated with class-imbalanced SSFER effectively. Extensive experiments on four facial expression datasets demonstrate the effectiveness of the proposed method across various imbalanced conditions. The source code is made publicly available at https://github.com/hangyu94/ReMoP.

1 Introduction

Facial expressions, a fundamental form of non-verbal communication, are crucial for human-to-human and human-computer interactions [Li and Deng, 2022]. Recently, some semi-supervised facial expression recognition (SSFER) algorithms [Florea *et al.*, 2020; Li *et al.*, 2022; Roy and Etemad, 2024] have been proposed to improve model performance by assigning pseudo-labels to a large number of confident unlabeled samples. A common assumption is that the class distribution of the constructed semi-supervised facial expression datasets is balanced, meaning the number of facial expression samples in each class is nearly equal. However, this assumption may be unsatisfactory in realistic scenarios, as the class distributions in most facial expression datasets are imbalanced [Li *et al.*, 2021; Zhang *et al.*, 2023].



(a) Generate Pseudo-labels (b) F1 score of pseudo-label predictions

Figure 1: (a) Illustration of generating pseudo-labels for unlabeled samples using a linear classifier and a feature classifier. (b) Comparison of F1 score of predictions between the linear pseudo-label, the feature-based pseudo-label, and their combination. We conduct experiments on RAF-DB under the setting of $(N_1, \gamma) = (2000, 50)$.

It is well known that the model trained on class-imbalanced labeled samples is biased towards the majority classes [Menon *et al.*, 2021], for example, Happiness and Neutral in FER. This issue can be further exacerbated for existing semi-supervised learning (SSL) methods, since the pseudo-labels assigned to confident unlabeled samples may also be biased, leading to an even more severely imbalanced training set. However, traditional class-imbalanced learning methods, usually designed for labeled samples, cannot be easily combined with SSL methods [Lee *et al.*, 2021]. Therefore, to better facilitate real-world scenarios, we aim to *design an SSFER method that generates high-quality pseudo-labels for imbalanced samples, which has not yet been thoroughly explored.*

Recently, some SSL methods [Lee *et al.*, 2021; Guo and Li, 2022; Ma *et al.*, 2024] combat the biased pseudo-labels caused by class imbalance at the level of linear prediction, but they often overlook the impact of feature similarity. This oversight becomes more exacerbated in FER, where the main challenge is *significant intra-class variations and inter-class similarities at the feature level* [Li *et al.*, 2019; Ruan *et al.*, 2021]. In this context, feature similarity and linear prediction are crucial for generating high-quality pseudo-labels. To this end, we first examine the effects of linear prediction and feature similarity on pseudo-label quality. There are three sources that guide the model to pseudo-label unlabeled samples: one from a *linear classifier*, one from a *feature classifier* (*e.g.*, a collection of class centers [Wen *et al.*,

[*] Corresponding author

2016]), and one from the combination of both classifiers. As shown in Figure 1(b), we observe that the mixed manner consistently enhances the quality of pseudo-labels with higher F1 scores than others. This observation provides excellent insights into learning unlabeled facial expression samples by considering feature and linear prediction levels.

In this paper, we propose a simple yet effective method for class-imbalanced SSFER with a regularized mixture of predictions, which performs a reliable way to generate pseudo-labels for imbalanced samples. Specifically, we first use a feature classifier to learn a *feature prediction, reflecting the feature similarity between a facial expression feature and the class centers* [Wen *et al.*, 2016]. Meanwhile, a linear classifier projects the feature to a linear prediction. We then merge the linear and feature ones into a mixture of predictions to generate a high-quality pseudo-label for an unlabeled sample. However, merely combining these two predictions cannot fully mitigate the bias from the class imbalance issue. To overcome this challenge, inspired by neural collapse [Papyan *et al.*, 2020], we introduce a class regularization term into the feature classifier that constrains the geometry of the feature distribution. This ensures the discriminative power of facial expression features with an explicit inter-class boundary. Furthermore, our method is effective in classifying facial expressions during the inference stage. Overall, the main contributions of this work can be summarized as follows:

- To the best of our knowledge, this is the first practical solution for class-imbalanced semi-supervised facial expression recognition. As a plug-and-play solution, it exhibits strong transferability across various SSL and SSFER methods like FixMatch and Ada-CM.
- We design a mixture of predictions to generate high-quality pseudo-labels and enhance inference performance. In addition, we introduce a class regularization term to improve the discriminative power of features.
- Extensive experiments on four challenging datasets demonstrate the superiority of our method under various imbalanced settings, compared to the state-of-the-art class-imbalanced SSL algorithms.

2 Related Work

Facial Expression Recognition. An underlying objective of Facial Expression Recognition (FER) is to extract discriminative facial expression features along with a linear classifier for predicted class distributions. In deep learning, numerous FER methods have been proposed under this paradigm [Li *et al.*, 2024]. Typically, Li *et al.* [2019] proposed classifying facial expressions with attention mechanism from partially-occluded faces. Wang *et al.* [2020] proposed a suppressing method to overcome the uncertainty issue in FER. Xue *et al.* [2021] applied the Vision Transformers to explore the relation-aware facial expression features. Wu *et al.* [2023] leveraged facial landmarks to reduce the impact of noisy supervision. Until recently, the issue of class imbalance has also been widely explored in FER. For example, Li *et al.* [2021] designed an adaptive regular loss to re-weight the importance of different facial expression categories. Zeng *et al.* [2022] leveraged large-scale unlabeled images to mitigate the data bias from class-imbalanced facial expression samples. Zhang *et al.* [2023] proposed to extract extra information related to the minority category from all training samples.

While most of these methods have achieved superior performance in a fully-supervised manner, they heavily consume a large number of labeled samples for the model's training. Unlike them, several methods have leveraged unlabeled samples to explore FER performance in a semi-supervised manner [Jiang and Deng, 2023]. Particularly, Florea *et al.* [2020] proposed to predict artificial labels of unlabeled samples by center embeddings. Li *et al.* [2022] designed an adaptive confidence margin to fully learn unlabeled samples. Du *et al.* [2023] further considered the label ambiguity issue among labeled samples and enhanced the learning on unlabeled samples. *However, all these methods are designed under the balanced class distribution. To the best of our knowledge, we are the first to achieve SSFER with class-imbalanced facial expression samples.*

Class-Imbalanced Semi-Supervised Learning. Recent semi-supervised learning (SSL) has a long history of research. A typical pipeline in SSL generates pseudo-labels for confident unlabeled samples using the model's outputs, which are used for supervised learning. Take the recent FixMatch [Sohn *et al.*, 2020] as an example. Specifically, it generated pseudo-labels for unlabeled samples with high-confidence predictions above a pre-defined threshold, then leveraged their weak and strong augmentations to achieve consistency regularization. While these methods have seen success in the balanced class scenarios, they fail to improve the biased model's performance when encountering class-imbalanced training samples.

To alleviate the above practical problem, class-imbalanced semi-supervised learning (CISSL) methods have garnered increasing attention in recent years [Oh *et al.*, 2022; Wei and Gan, 2023; Lee and Kim, 2024]. For example, Kim *et al.* [2020] proposed softly refining the biased pseudo-labels by solving a convex optimization problem. Wei *et al.* [2021] selected more pseudo-labeled samples from minority categories to retrain the baseline SSL model. Lee *et al.* [2021] introduced an auxiliary balanced classifier in FixMatch to mitigate the class imbalance. Unlike the fixed confidence thresholding, Guo *et al.* [2022] selected pseudo-labeled samples based on the adaptive thresholds for different categories. Yu *et al.* [2023] introduced the concept of energy scores from out-of-distribution detection to generate pseudo-labels for unlabeled samples, which can address the drawbacks of previous confidence scores. Very recently, Ma *et al.* [2024] proposed to model various class distributions by multiple experts trained with different logit adjustments. *These methods mainly focus on the predictions from linear classifiers but ignore the importance of feature similarity. In contrast, we discover the solution of a regularized mixture of predictions for pseudo-labeling class-imbalanced unlabeled samples.*

3 Preliminary

The Neural Collapse ($\mathcal{NC}$) [Papyan *et al.*, 2020] reveals an intriguing phenomenon that during the terminal phase of train-

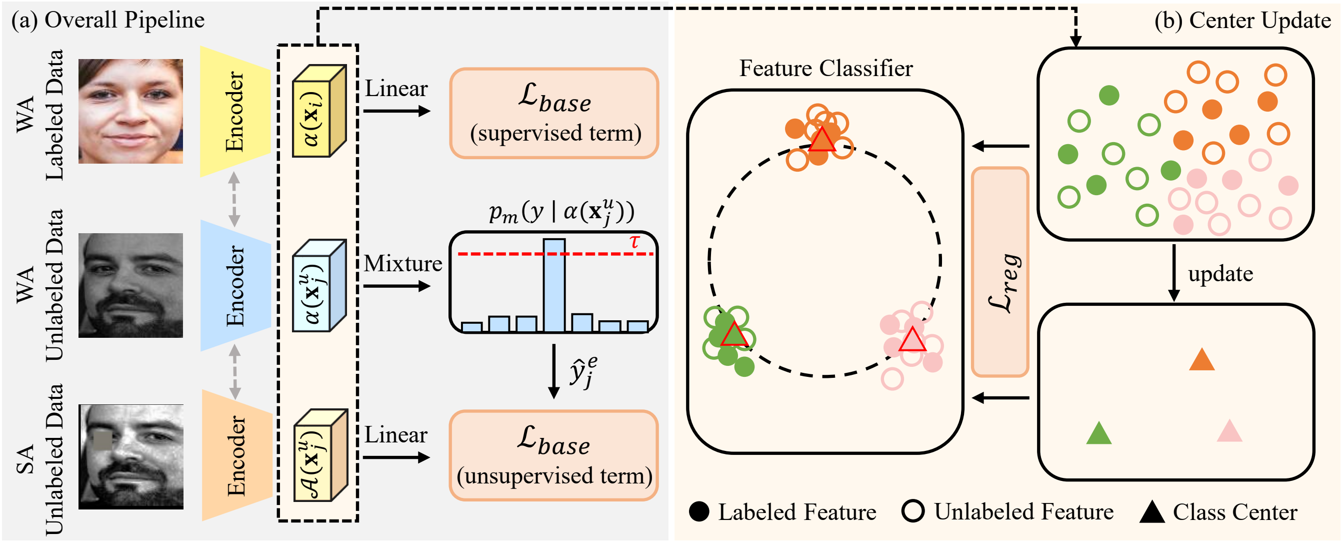


Figure 2: Illustration of the proposed ReMoP method. (a) The gray square part represents the overall pipeline with labeled and unlabeled samples. (b) The yellow square part is a class regularization branch to make class centers collapse to a simplex ETF structure during training. For clarity, we present the weight-shared encoders with three colors to distinguish different inputs.

ing a deep classification model on the balanced dataset, the learned features of the same class will converge to their class centers. Meanwhile, these class centers will exhibit a simplex equiangular tight frame (ETF). The ETF structure can maximize the inter-class difference.

Definition 1 (Simplex Equiangular Tight Frame) For a K-class classification problem, a simplex ETF is a collection of vectors $\mathbf{m}_k \in \mathbb{R}^d$, $k = 1, 2, ..., K$ if:

$$\mathbf{M} = \sqrt{\frac{K}{K-1}}\mathbf{U}\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right), \tag{1}$$

where $\mathbf{M} = [\mathbf{m}_1, ..., \mathbf{m}_K] \in \mathbb{R}^{d\times K}$, d is the dimension of the vector, $\mathbf{U} \in \mathbb{R}^{d\times K}$ is an orthogonal matrix satisfying $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_K$, $\mathbf{I}_K$ is an identity matrix, and $\mathbf{1}_K \in \mathbb{R}^K$ is a vector of all ones.

Then, the vectors in a simplex ETF construct an optimal geometry for image classification, which have an equal l_2 norm and the same pair-wise angle, *i.e.*,

$$\mathbf{m}_k^\top\mathbf{m}_t = \frac{K}{K-1}\delta_{k,t} - \frac{1}{K-1}, \forall k, t \in \{1, 2, ..., K\}, \tag{2}$$

where $\delta_{k,t}$ equals to 1 when $k = t$ and 0 otherwise. The pair-wise angle $-\frac{1}{K-1}$ is the maximal equiangular separation of K vectors in the feature space $\mathbb{R}^d$ [Yang *et al.*, 2022].

Based on the above definition, two important geometry properties derived from the neural collapse phenomenon can be formally summarized as: ($\mathcal{NC}_1$) **Variability collapse.** Intra-class variability of the last-layer features collapses to zero during the terminal phase of training, *i.e.*, $||\mathbf{x}_i - \mathbf{h}_k|| = 0$, where $\mathbf{x}_i$ is i-th feature from class k, and $\mathbf{h}_k$ is the center of class k; ($\mathcal{NC}_2$) **Convergence to a simplex ETF.** The normalized class centers, *i.e.*, $\widetilde{\mathbf{h}}_k = (\mathbf{h}_k - \mathbf{h}_G)/||\mathbf{h}_k - \mathbf{h}_G||$, converge to a simplex ETF satisfying Eq. (2), where $\mathbf{h}_G = \text{Avg}_k\{\mathbf{h}_k\}$ is the global mean of all centers.

4 Method

4.1 Problem Formulation

For a K-class imbalanced semi-supervised FER task, we have both a batch of labeled samples $\mathcal{X} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ and unlabeled samples $\mathcal{U} = \{\mathbf{x}_j^u\}_{j=1}^{\mu N}$, where μ is a hyper-parameter to determine the relative ratio of $\mathcal{X}$ and $\mathcal{U}$, and $\mathbf{x}_i$, $\mathbf{x}_j^u \in \mathbb{R}^d$ are facial expression features for labeled and unlabeled samples, respectively. For the i-th labeled feature $\mathbf{x}_i$, it is associated with a ground-truth label $y_i \in \{1, 2, ..., K\}$. Let N_k denote the number of labeled samples in class k, we assume that the classes are sorted in descending order, *i.e.* $N_1 \geq N_2 \geq ... \geq N_K$. The degree of class imbalance is defined as the imbalance ratio $\gamma = \frac{N_1}{N_K}$. Similarly, the same assumption for unlabeled samples exists in the CISSL setting. Considering that the training samples from existing facial expression datasets are inherently class-imbalanced [Zhang *et al.*, 2023], we therefore preserve the natural effect of class imbalance on unlabeled samples.

A basic goal of class-imbalanced semi-supervised FER is to train a linear classifier $\mathcal{F}_l$ parameterized by θ_l using $\mathcal{X}$ and $\mathcal{U}$. Specifically, let $p_l(y \mid \mathbf{x}) = \mathcal{F}_l(\mathbf{x}; \theta_l)$ be the linear prediction produced by the linear classifier for input $\mathbf{x}$, there is an objective similar to FixMatch [Sohn *et al.*, 2020], consisting of a supervised term and an unsupervised term:

$$\begin{aligned}\mathcal{L}_{base} = &\underbrace{\sum_{i=1}^{N}\mathcal{H}\left(y_i, p_l\left(y \mid \alpha\left(\mathbf{x}_i\right)\right)\right)}_{\text{supervised}} \\ &+ \underbrace{\sum_{j=1}^{\mu N}\mathbb{1}(\max(\mathbf{y}_j) \geq \tau)\mathcal{H}(\widehat{y}_j, p_l(y \mid \mathcal{A}(\mathbf{x}_j^u)))}_{\text{unsupervised}},\end{aligned} \tag{3}$$

where $\mathcal{H}$ is the cross-entropy loss, $\mathbb{1}(\cdot)$ is the indicator function, and τ is a threshold to decide whether or not to retain the pseudo-label. $\alpha(\cdot)$ and $\mathcal{A}(\cdot)$ denote the weakly-augmented operation and strongly-augmented operation, respectively. To obtain the pseudo-label $\widehat{y}_j = \underset{k}{\text{argmax}}(\mathbf{y}_j)$, it is general to compute the model's prediction for the weakly-augmented version of the unlabeled sample: $\mathbf{y}_j = p_l(y \mid \alpha(\mathbf{x}_j^u))$[1]. In this regard, our work aims to enhance pseudo-label quality at the feature level.

4.2 The Proposed Method

In class-imbalanced semi-supervised facial expression recognition, the core challenge is *how to generate high-quality pseudo-labels for imbalanced facial expression samples*. To address this issue, we propose a novel method, named **Re**gularized **M**ixture **o**f **P**redictions (ReMoP). In the following section, we will provide an overview of the proposed method and elaborate on key technologies.

Overview. Figure 2 depicts the framework of our proposed ReMoP method. Specifically, in each forward pass, a labeled sample and an unlabeled sample are input into the same encoder for their features using weakly-augmented (WA) and strongly-augmented (SA) operations. Then, we follow the standard SSL learner to project the WA labeled feature and the SA unlabeled feature to the linear predictions using a linear classifier. To generate high-quality pseudo-labels for unlabeled samples, we project the WA unlabeled feature to a mixture of predictions $p_m(y \mid \alpha(\mathbf{x}_j^u))$, which combines a feature prediction from the feature classifier and a linear prediction from the linear classifier. Meanwhile, we select all labeled features and some unlabeled features with confidence scores higher than the threshold τ, to update the feature classifier. Finally, a class regularization term $\mathcal{L}_{reg}$ is introduced to constrain the geometry of the feature distribution.

Learning a mixture of predictions. In this work, we combine the linear prediction and the feature prediction (*i.e.*, feature similarity). Instead of linear pseudo-labels in the existing SSL learner, we use a mixture of predictions to generate high-quality pseudo-labels for unlabeled samples. Specifically, given a linear classifier $\mathcal{F}_l$, we first learn a linear prediction $p_l(y \mid \alpha(\mathbf{x}_j^u))$ for a WA unlabeled feature $\alpha(\mathbf{x}_j^u)$. To execute the mixture of predictions, we initialize a feature classifier $\mathbf{H} \in \mathbb{R}^{d \times K}$ by

$$\mathbf{H} = [\mathbf{h}_1, ..., \mathbf{h}_K], \tag{4}$$

where $\mathbf{h}_k \in \mathbb{R}^{d \times 1}$ denotes the center for the k-th class. Let $p_s(y \mid \alpha(\mathbf{x}_j^u))$ be the feature prediction produced by the feature classifier, we normalize the feature and the k-th center $\mathbf{h}_k$ for their similarity score as

$$p_s^k(y \mid \alpha(\mathbf{x}_j^u)) = \frac{\alpha(\mathbf{x}_j^u) \cdot \mathbf{h}_k}{||\alpha(\mathbf{x}_j^u)||||\mathbf{h}_k||}. \tag{5}$$

Then, we combine the above two predictions and obtain a mixture of predictions as

$$p_m(y \mid \alpha(\mathbf{x}_j^u)) = \beta p_l + (1-\beta) p_s, \tag{6}$$

where β is a hyper-parameter to balance two predictions. Finally, we use the mixture of predictions to obtain an enhanced pseudo-label:

$$\widehat{y}_j^e = \underset{k}{\text{argmax}}(p_m(y \mid \alpha(\mathbf{x}_j^u))), \tag{7}$$

which is used to replace the linear pseudo-label for the unsupervised term in Eq. (3).

Algorithm 1 ReMoP's main learning algorithm.

```
Input: A batch of labeled data $\mathcal{X} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ and unlabeled
    data $\mathcal{U} = \{\mathbf{x}_j^u\}_{j=1}^{\mu N}$, number of epochs $E_{max}$, number of train-
    ing iterations $t_{max}$, and the initial parameters $\theta_l$ and $\mathbf{H}$.
Output: Updated parameters $\theta_l$ and $\mathbf{H}$.
// Training
for $e = 1, 2, ..., E_{max}$ do
    for $t = 1, 2, ..., t_{max}$ do
        // Exploring labeled and unlabeled samples
        Learn the linear prediction $p_l(y \mid \alpha(\mathbf{x}_i))$ for labeled sample $\mathbf{x}_i$ using $\mathcal{F}_l$.
        Learn the mixture of predictions $p_m(y \mid \alpha(\mathbf{x}_j^u))$ for unlabeled sample $\mathbf{x}_j^u$ by Eq. (6).
        Generate the pseudo-label $\widehat{y}_j^e$ by Eq. (7).
        Update model parameters by Eq. (3).
        // Learning a feature classifier
        Update class centers using labeled features and confident unlabeled features by $\mathcal{L}_c$.
        Constrain the geometry of centers by Eq. (8).
    end for
end for
// Testing
Deploy classifiers $\mathcal{F}_l$ and $\mathbf{H}$ for the mixture of predictions.
```

Class regularization term. As previously mentioned, the learning behavior of a classification model on balanced datasets is revealed by a neural collapse phenomenon [Papyan *et al.*, 2020]. To be specific, when the model achieves zero training error rate, the learned features will collapse to their class centers, forming an interesting geometric structure (*i.e.*, a simplex ETF) after being globally centered.

However, the model cannot exhibit the ideal structure on imbalanced datasets, resulting in the features of minority classes becoming indistinguishable [Fang *et al.*, 2021]. To this end, we claim that the cosine similarity of any pairs of globally-centered vectors from $\mathbf{H}$ should be close to $-1/(K-1)$ in the imbalanced setting. In this way, we can maximally separate inter-class features. Specifically, we first update class centers [Wen *et al.*, 2016] using labeled samples and confident unlabeled samples via the center loss $\mathcal{L}_c = \frac{1}{2}\sum ||\mathbf{x} - \mathbf{h}_y||_2^2$, where $\mathbf{h}_y$ denotes the y-th center related to the feature $\mathbf{x}$. Then, we introduce a regularization term $\mathcal{L}_{reg}$ into the mixture of predictions to further mitigate the data bias from the class imbalance:

$$\mathcal{L}_{reg} = \sum_{k,t=1}^{K} \mathbb{1}(k \neq t) \left(\widetilde{\mathbf{h}}_k^\top \widetilde{\mathbf{h}}_t - \left(-\frac{1}{K-1} \right) \right)^2. \tag{8}$$

In summary, our method is optimized in an end-to-end process. It is flexible to be integrated into any standard SSL and SSFER frameworks. Finally, we have the following loss function for training:

$$\mathcal{L}_{total} = \mathcal{L}_{base} + \lambda_1 \mathcal{L}_c + \lambda_2 \mathcal{L}_{reg}, \tag{9}$$

[1] In SSL, it is popular to generate a pseudo-label using the prediction on the weakly-augmented unlabeled sample, which is used to match the prediction on the strongly-augmented version.

where λ_1 and λ_2 are hyper-parameters to balance three terms' intensity. Note that $\mathcal{L}_{base}$ takes the enhanced pseudo-labels in Eq. (7) as inputs.

Moreover, unlike the linear prediction from a linear classifier used in existing FER pipelines, we merge it to the feature similarity for a mixture of predictions during the testing stage. Specifically, for a testing feature $\mathbf{x}^t$, we deploy two classifiers $\mathcal{F}_l$ and $\mathbf{H}$ for the mixture of predictions $p_m(y \mid \mathbf{x}^t)$. The whole progress of the proposed method is summarized in Algorithm 1.

5 Experiments

5.1 Datasets

We conduct experiments on four public facial expression datasets, including RAF-DB [Li and Deng, 2019], FERPlus [Barsoum *et al.*, 2016], CK+ [Lucey *et al.*, 2010], and AffectNet [Mollahosseini *et al.*, 2017]. **RAF-DB** is a large-scale facial expression dataset with 29,672 real-world facial images, which are labeled by about 40 annotators into a single-label subset and a two-tab subset. In our experiments, we use the single-label subset with 12,271 training images and 3,068 testing images, including seven basic facial expression categories (*i.e.*, surprise, fear, disgust, happiness, sadness, anger, and neutral). **FERPlus** provides the new eight-class labels (*i.e.*, seven basic categories and contempt) created by 10 crowd-sourced annotators. It consists of 28,709 training images, 3,589 validation images, and 3,589 testing images. **CK+** consists of 593 video sequences from 123 subjects. For a fair comparison, we follow [Li *et al.*, 2022] to select each sequence's first frame and the last frame as the neutral face and the targeted facial expression, containing 636 facial images with seven basic categories. Different from the first three datasets, **AffectNet** is by far the largest in-the-wild facial expression dataset, containing more than 1M facial images but a large number of noisy labels. In our experiments, we choose about 287,651 manually annotated images with seven basic categories and the contempt category as inter-dataset unlabeled samples to evaluate the CISSL performance.

Following the setting in CISSL [Guo and Li, 2022], we use the imbalance ratio γ with the given N_1 to construct the class-imbalanced training set. Specifically, we set the number of labeled samples in class k as $N_k = N_1 \cdot \gamma^{-\frac{k-1}{K-1}}$ for $1 < k \leq K$. In this work, we design various combinations of γ and N_1. Considering the imbalanced testing class distribution in RAF-DB, FERPlus, and CK+, we report the overall accuracy and the mean accuracy across all categories by default. Unless otherwise specified, we conduct experiments three times using different random seeds to obtain the mean and the standard deviation.

Method	Strategy		RAF-DB	FERPlus	CK+
	$\mathbf{H}$	$\mathcal{L}_{reg}$	$N_1 = 500$ $\gamma = 150$	$N_1 = 1000$ $\gamma = 150$	$N_1 = 500$ $\gamma = 150$
Vanilla	-	-	73.50/ 49.10	76.75/ 49.51	77.36/ 55.42
	✓	-	74.22/ 50.84	77.74/ 52.87	78.30/ 57.05
	✓	✓	75.13/ 52.16	78.08/ 53.24	79.09/ 57.67
FixMatch	-	-	77.71/ 55.48	82.21/ 54.62	80.03/ 59.06
	✓	-	79.86/ 58.52	82.65/ 57.16	84.43/ 63.51
	✓	✓	81.29/ 60.93	83.67/ 58.77	85.69/ 66.69

Table 1: Ablation study of different modules in our method on RAF-DB, FERPlus, and CK+ (in %, overall/ mean accuracy). We conduct experiments over the random seed as 1. Vanilla denotes that the model is trained using limited labeled samples. This also applies to the following tables. Note that the performance on the CK+ dataset is reported by training on RAF-DB and evaluating on CK+.

5.2 Implementation Details

We implement all experiments using the PyTorch toolbox with one NVIDIA A100 GPU. For the basic encoder, we use the ResNet-18 [He *et al.*, 2016] pre-trained on the MS-Celeb-1M face recognition dataset [Guo *et al.*, 2016] for learning facial expression features. Besides, we use MTCNN [Zhang *et al.*, 2016] to align and resize facial images to 224×224 pixels. During training, we follow [Li *et al.*, 2022] to use *RandomCrop* and *RandomHorizontalFlip* as the weak augmentation strategy, and add RandAugment [Cubuk *et al.*, 2020] as the strong augmentation strategy for a fair comparison. By default, we use the Adam optimizer. For RAF-DB and CK+, we train the model with a learning rate of $1e-4$, training epoch 100, and batch size 16. For FERPlus, due to a large number of training samples, we train the model with a learning rate of $1e-4$, training epoch 80, and batch size 32. The number of training iterations t_{max} is set to 1,000 in all experiments. The relative ratio μ is set to 1 except for AffectNet as 5. The hyper-parameter β in Eq. (6) is set to 0.5. Following FixMatch [Sohn *et al.*, 2020], the default threshold τ is 0.95. In Eq. (9), the hyper-parameters λ_1 and λ_2 are set to $1e-4$ and 0.1, respectively.

5.3 Ablation Study

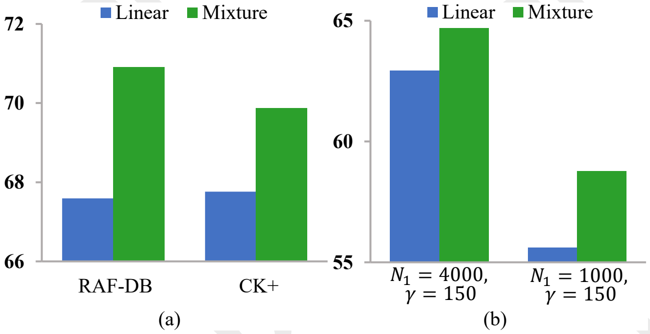
Effect of two modules in ReMoP. We examine the effectiveness of ReMoP in Table 1. Several observations can be summarized as follows: 1) Compared with the Vanilla-based baseline using a linear classifier (row 1), introducing a feature classifier for the mixture of predictions (row 2) consistently improves inference performance; 2) A significant improvement between rows 2 and 3 is achieved after introducing $\mathcal{L}_{reg}$ into $\mathbf{H}$. Since the class regularization term is used to constrain the feature classifier for the geometry of feature distribution, it is reasonable to enhance the discriminative power of facial expression features; 3) The similar improvements (rows 4 to 6) are present in the learning of unlabeled samples as well. These results also verify that the proposed method effectively learns unlabeled facial expression samples at feature and linear prediction levels.

Effect of the mixture of predictions for testing samples. To eliminate the concern that the mixture of predictions is more effective in determining facial expression categories than the linear prediction during the testing stage, we conduct an ablation study to evaluate their difference. As shown in Figure 3, the mixture strategy consistently improves performance in each case. This contributes an important insight for FER to explore the impact of discriminative features.

Effect of varying hyper-parameter β. In Figure 4, we conduct an ablation study on RAF-DB to analyze the effect of different value $\beta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The hyper-parameter β reflects the intensity of two predictions in the

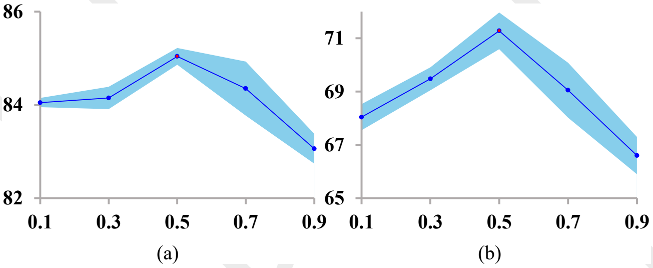
Method	RAF-DB		CK+	
	$N_1 = 2000$ $\gamma = 50$	$N_1 = 500$ $\gamma = 50$	$N_1 = 2000$ $\gamma = 150$	$N_1 = 500$ $\gamma = 150$
FixMatch [Sohn *et al.*, 2020]	$85.31_{\pm 0.21}$/ $72.19_{\pm 0.54}$	$80.99_{\pm 0.33}$/ $64.46_{\pm 1.05}$	$83.18_{\pm 1.34}$/ $67.76_{\pm 0.67}$	$79.93_{\pm 2.05}$/ $57.92_{\pm 4.86}$
w/ ReMoP	$86.85_{\pm 0.07}$/ $76.92_{\pm 0.08}$	$83.15_{\pm 0.60}$/ $71.12_{\pm 0.02}$	$85.06_{\pm 0.68}$/ $69.10_{\pm 0.90}$	$84.70_{\pm 0.77}$/ $65.90_{\pm 0.65}$
Ada-CM [Li *et al.*, 2022]	$85.75_{\pm 0.29}$/ $73.12_{\pm 0.39}$	$81.39_{\pm 0.36}$/ $65.80_{\pm 1.49}$	$83.80_{\pm 1.10}$/ $68.05_{\pm 1.29}$	$80.24_{\pm 1.49}$/ $60.26_{\pm 1.54}$
w/ ReMoP	$86.15_{\pm 0.36}$/ $76.01_{\pm 0.69}$	$81.63_{\pm 0.47}$/ $68.80_{\pm 0.27}$	$84.05_{\pm 0.23}$/ $69.99_{\pm 0.11}$	$83.73_{\pm 1.81}$/ $67.12_{\pm 0.77}$

Table 2: Ablation study in terms of overall/ mean accuracy using different SSL and SSFER learners on RAF-DB and CK+ (in %, mean ± standard deviation).



(a) (b)

Figure 3: The effect of the mixture of predictions during the testing stage in terms of mean accuracy on (a) RAF-DB and CK+ under the setting of $(N_1, \gamma) = (2000, 150)$, (b) FERPlus. We conduct experiments over the random seed as 1.



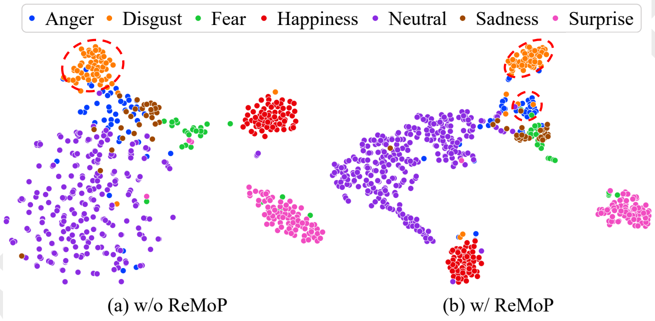
(a) (b)

Figure 4: The effect of hyper-parameter β on RAF-DB in terms of (a) overall and (b) mean accuracy under the setting of $(N_1, \gamma) = (2000, 150)$.

mixture of predictions. From the results, we can observe that when β is too large or too small, the performance decreases as the excessive importance of linear prediction or feature similarity hampers the quality of pseudo-labels. We fix $\beta = 0.5$ for all the experiments according to the experimental results.

Effect of ReMoP on feature learning. As mentioned above, the proposed ReMoP constrains the geometry of the feature distribution. To verify this, we visualize the t-SNE distribution [Van der Maaten and Hinton, 2008] for facial expression features learned by FixMatch and FixMatch-based ReMoP. As shown in Figure 5, ReMoP helps construct a discriminative feature space, which can boost FER performance.

Effect of ReMoP using different SSL learners. In this work, we claim that the proposed method as a plug-and-play solution can be integrated into existing SSL and SSFER methods. To evaluate this, we conduct experiments using a typical SSL method (FixMatch) and an SSFER method (Ada-CM). As shown in Table 2, using ReMoP further improves the performance on overall and mean accuracy in each case, demonstrating the ability of the proposed method to address the class-imbalanced issue in SSFER. In the following, we use FixMatch [Sohn *et al.*, 2020] as the baseline SSL learner.



(a) w/o ReMoP (b) w/ ReMoP

Figure 5: 2D t-SNE visualization of facial expression features from the CK+ dataset.

5.4 Comparison with State-of-the-Art Methods

In this section, based on the widely-used FixMatch method [Sohn *et al.*, 2020], we compare the proposed method with some popular CISSL methods, including ABC [Lee *et al.*, 2021], Adsh [Guo and Li, 2022], InPL [Yu *et al.*, 2023], and CPE [Ma *et al.*, 2024].

Intra-dataset evaluation. In this part, we assume that labeled and unlabeled samples come from the same dataset, that is, they are randomly divided from original training set.

As shown in Table 3, we compare the proposed ReMoP with state-of-the-art CISSL methods on RAF-DB under four different class distributions. From these results, we can see that 1) the proposed method can consistently improve the vanilla method, validating the effectiveness of ReMoP to leverage unlabeled samples for discriminative learning; 2) the proposed method achieves satisfactory performance, outperforming existing methods by a large margin in most cases. For example, given the setting of $(N_1, \gamma) = (2000, 150)$, the proposed ReMoP significantly outperforms the competitive Adsh [Guo and Li, 2022] by 1.80% and 3.34% on overall and average accuracy, respectively.

In addition, Table 4 compares the performance on CK+ and FERPlus under two different class distributions, respectively. We can see that the proposed ReMoP steadily and

Method	RAF-DB			
	$N_1 = 2000$ $\gamma = 50$	$N_1 = 2000$ $\gamma = 150$	$N_1 = 500$ $\gamma = 50$	$N_1 = 500$ $\gamma = 150$
Vanilla	84.04$_{\pm 0.56}$/ 70.13$_{\pm 0.73}$	80.97$_{\pm 0.30}$/ 61.41$_{\pm 1.05}$	77.55$_{\pm 0.40}$/ 58.28$_{\pm 1.15}$	73.08$_{\pm 0.39}$/ 48.52$_{\pm 1.54}$
FixMatch [Sohn *et al.*, 2020]	85.31$_{\pm 0.21}$/ 72.19$_{\pm 0.54}$	82.39$_{\pm 0.52}$/ 65.15$_{\pm 1.19}$	80.99$_{\pm 0.33}$/ 64.46$_{\pm 1.05}$	77.33$_{\pm 0.27}$/ 53.88$_{\pm 1.13}$
w/ ABC [Lee *et al.*, 2021]	85.93$_{\pm 0.33}$/ 73.85$_{\pm 0.11}$	83.25$_{\pm 0.52}$/ 68.40$_{\pm 0.94}$	82.63$_{\pm 0.03}$/ 68.07$_{\pm 0.76}$	78.90$_{\pm 0.53}$/ 58.78$_{\pm 1.34}$
w/ Adsh [Guo and Li, 2022]	86.15$_{\pm 0.54}$/ 74.19$_{\pm 1.30}$	83.24$_{\pm 0.59}$/ 67.94$_{\pm 1.33}$	82.55$_{\pm 0.16}$/ 68.22$_{\pm 1.16}$	80.12$_{\pm 0.31}$/ 62.52$_{\pm 0.85}$
w/ InPL [Yu *et al.*, 2023]	85.86$_{\pm 0.16}$/ 73.44$_{\pm 0.21}$	82.94$_{\pm 0.27}$/ 67.31$_{\pm 0.49}$	82.19$_{\pm 0.15}$/ 67.41$_{\pm 0.94}$	78.74$_{\pm 0.23}$/ 58.28$_{\pm 1.63}$
w/ CPE [Ma *et al.*, 2024]	85.40$_{\pm 0.24}$/ **78.21**$_{\pm 0.23}$	83.80$_{\pm 0.65}$/ **73.02**$_{\pm 0.23}$	80.66$_{\pm 0.43}$/ 70.61$_{\pm 0.01}$	78.25$_{\pm 0.35}$/ 62.90$_{\pm 1.86}$
w/ **ReMoP (Ours)**	**86.85**$_{\pm 0.07}$/ 76.92$_{\pm 0.08}$	**85.04**$_{\pm 0.18}$/ 71.28$_{\pm 0.69}$	**83.15**$_{\pm 0.60}$/ **71.12**$_{\pm 0.02}$	**81.42**$_{\pm 0.36}$/ **64.89**$_{\pm 2.84}$

Table 3: Performance comparison in terms of overall/ mean accuracy with the state-of-the-art CISSL methods on RAF-DB (in %, mean ± standard deviation). The best results are shown in **bold**. This also applies to the following tables.

Method	CK+		FERPlus	
	$N_1 = 2000$ $\gamma = 150$	$N_1 = 500$ $\gamma = 150$	$N_1 = 4000$ $\gamma = 150$	$N_1 = 1000$ $\gamma = 150$
Vanilla	82.44$_{\pm 0.32}$/ 63.53$_{\pm 1.42}$	76.94$_{\pm 1.44}$/ 54.34$_{\pm 0.78}$	82.79$_{\pm 0.32}$/ 60.14$_{\pm 0.73}$	77.13$_{\pm 0.33}$/ 49.44$_{\pm 0.79}$
FixMatch [Sohn *et al.*, 2020]	83.18$_{\pm 1.34}$/ 67.76$_{\pm 0.67}$	79.93$_{\pm 2.05}$/ 57.92$_{\pm 4.86}$	84.70$_{\pm 0.30}$/ 63.62$_{\pm 1.05}$	82.02$_{\pm 0.14}$/ 54.26$_{\pm 0.67}$
w/ ABC [Lee *et al.*, 2021]	83.96$_{\pm 0.66}$/ 68.69$_{\pm 0.88}$	80.08$_{\pm 2.55}$/ 59.45$_{\pm 3.79}$	84.99$_{\pm 0.22}$/ 63.33$_{\pm 0.22}$	82.47$_{\pm 0.43}$/ 56.16$_{\pm 0.11}$
w/ Adsh [Guo and Li, 2022]	83.65$_{\pm 1.23}$/ 66.37$_{\pm 2.27}$	83.17$_{\pm 1.52}$/ 65.18$_{\pm 1.24}$	84.89$_{\pm 0.01}$/ 63.31$_{\pm 0.87}$	81.95$_{\pm 0.22}$/ 56.47$_{\pm 0.65}$
w/ InPL [Yu *et al.*, 2023]	83.55$_{\pm 0.73}$/ 66.67$_{\pm 0.52}$	81.92$_{\pm 2.02}$/ 61.62$_{\pm 3.53}$	84.41$_{\pm 0.32}$/ 63.00$_{\pm 0.92}$	81.58$_{\pm 0.09}$/ 55.21$_{\pm 0.69}$
w/ CPE [Ma *et al.*, 2024]	82.63$_{\pm 0.39}$/ 68.27$_{\pm 0.86}$	80.27$_{\pm 2.59}$/ 64.20$_{\pm 0.43}$	83.19$_{\pm 0.63}$/ 64.25$_{\pm 0.60}$	79.86$_{\pm 0.23}$/ 56.88$_{\pm 0.48}$
w/ **ReMoP (Ours)**	**85.06**$_{\pm 0.68}$/ **69.10**$_{\pm 0.90}$	**84.70**$_{\pm 0.77}$/ **65.90**$_{\pm 0.65}$	**85.81**$_{\pm 0.17}$/ **65.01**$_{\pm 0.96}$	**83.03**$_{\pm 0.46}$/ **57.14**$_{\pm 1.20}$

Table 4: Performance comparison in terms of overall/ mean accuracy with the state-of-the-art CISSL methods on CK+ and FERPlus (in %, mean ± standard deviation).

Method	RAF-DB	FERPlus
Baseline*	87.42	86.06
RUL* [2021]	88.98	88.30
MEK* [2023]	89.77	-
Pseudo-Labeling [2013]	87.39	85.13
Mean-Teacher [2017]	88.41	86.15
PT [2023]	88.69	86.60
Ada-CM [2022]	89.28	87.80
FixMatch [2020]	87.74	86.45
w/ ABC [2021]	87.78	87.38
w/ Adsh [2022]	88.98	88.01
w/ InPL [2023]	89.54	88.23
w/ CPE [2024]	88.14	87.98
w/ **ReMoP (Ours)**	**90.29**	**89.06**

Table 5: Performance comparison in terms of overall accuracy on RAF-DB and FERPlus using AffectNet as inter-dataset unlabeled samples. We conduct experiments over one random seed. *The model is trained using original training samples without extra unlabeled samples.

significantly outperforms the state-of-the-art CISSL methods. For example, ReMoP outperforms CPE [Ma *et al.*, 2024] by 2.62% and 3.17% overall accuracy on FERPlus under two different settings, respectively. These results demonstrate the superiority of the proposed method in learning with class-imbalanced samples.

Inter-dataset evaluation. In this part, we assume that labeled and unlabeled samples come from different datasets.

In Table 5, we conduct experiments on RAF-DB and FERPlus. Following the setting [Jiang and Deng, 2023], we treat *the training images along with their labels from RAF-DB or FERPlus as labeled samples, and the training images from AffectNet as unlabeled samples.* From the table, we can see that the proposed ReMoP significantly outperforms the existing methods by a large margin. Specifically, on the RAF-DB dataset, our method achieves 2.15% overall accuracy improvements compared to the competitive CPE [Ma *et al.*, 2024] and outperforms PT [Jiang and Deng, 2023] by 1.6% overall accuracy. Furthermore, we compare the proposed ReMoP with some fully-supervised methods. For example, our method outperforms MEK [Zhang *et al.*, 2023] by 0.52% overall accuracy on RAF-DB.

6 Conclusion

In this work, we address the practical class-imbalanced semi-supervised facial expression recognition task, focusing on the quality of pseudo-labels for imbalanced samples in a reliable manner. To this end, we propose the Regularized Mixture of Predictions (ReMoP) method, which integrates feature similarity with linear prediction to create a mixture of predictions. Additionally, we introduce a class regularization term that effectively handles class-imbalanced feature distribution. Through extensive experiments on four facial expression datasets, we demonstrate the efficacy of ReMoP across various challenging imbalanced setups. Our work may further contribute to understanding the importance of feature similarity in classifying facial expressions.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants U22A2096, 62036007, 62376235, and 62306178, in part by Scientific and Technological Innovation Teams in Shaanxi Province under grant 2025RS-CXTD-011, in part by the Shaanxi Province Core Technology Research and Development Project under grant 2024QY2-GJHX-11, in part by the Fundamental Research Funds for the Central Universities under GrantQTZX23042, in part by the GDST Basic Research Fund Nos. 2022A1515011652 and 2024A1515012399, in part by the HKBU Faculty Niche Research Areas No. RC-FNRA-IG/22-23/SCI/04, and in part by the HKBU CSD Departmental Incentive Scheme.

Contribution Statement

Hangyu Li and Yixin Zhang contributed equally to this work.

References

[Barsoum *et al.*, 2016] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ICMI*, pages 279–283, 2016.

[Cubuk *et al.*, 2020] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR Workshops*, pages 702–703, 2020.

[Du *et al.*, 2023] Zhongjing Du, Xu Jiang, Peng Wang, Qizheng Zhou, Xi Wu, Jiliu Zhou, and Yan Wang. Lion: label disambiguation for semi-supervised facial expression recognition with progressive negative learning. In *IJCAI*, pages 699–707, 2023.

[Fang *et al.*, 2021] Cong Fang, Hangfeng He, Qi Long, and Weijie J. Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43):e2103091118, 2021.

[Florea *et al.*, 2020] Corneliu Florea, Mihai Badea, Laura Florea, Andrei Racoviteanu, and Constantin Vertan. Margin-mix: Semi-supervised learning for face expression recognition. In *ECCV*, pages 1–17, 2020.

[Guo and Li, 2022] Lanzhe Guo and Yufeng Li. Class-imbalanced semi-supervised learning with adaptive thresholding. In *ICML*, pages 8082–8094, 2022.

[Guo *et al.*, 2016] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, pages 87–102, 2016.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Jiang and Deng, 2023] Jing Jiang and Weihong Deng. Boosting facial expression recognition by a semi-supervised progressive teacher. *IEEE Transactions on Affective Computing*, 14(3):2402–2414, 2023.

[Kim *et al.*, 2020] Jaehyung Kim, Youngbum Hur, Sejun Park, Eunho Yang, Sung Ju Hwang, and Jinwoo Shin. Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. In *NeurIPS*, pages 14567–14579, 2020.

[Lee and Kim, 2024] Hyuck Lee and Heeyoung Kim. Cdmad: Class-distribution-mismatch-aware debiasing for class-imbalanced semi-supervised learning. In *CVPR*, pages 23891–23900, 2024.

[Lee *et al.*, 2021] Hyuck Lee, Seungjae Shin, and Heeyoung Kim. Abc: Auxiliary balanced classifier for class-imbalanced semi-supervised learning. In *NeurIPS*, pages 7082–7094, 2021.

[Lee, 2013] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*, page 896, 2013.

[Li and Deng, 2019] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2019.

[Li and Deng, 2022] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3):1195–1215, 2022.

[Li *et al.*, 2019] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2019.

[Li *et al.*, 2021] Hangyu Li, Nannan Wang, Xinpeng Ding, Xi Yang, and Xinbo Gao. Adaptively learning facial expression representation via cf labels and distillation. *IEEE Transactions on Image Processing*, 30:2016–2028, 2021.

[Li *et al.*, 2022] Hangyu Li, Nannan Wang, Xi Yang, Xiaoyu Wang, and Xinbo Gao. Towards semi-supervised deep facial expression recognition with an adaptive confidence margin. In *CVPR*, pages 4166–4175, 2022.

[Li *et al.*, 2024] Hangyu Li, Nannan Wang, Xi Yang, Xiaoyu Wang, and Xinbo Gao. Unconstrained facial expression recognition with no-reference de-elements learning. *IEEE Transactions on Affective Computing*, 15(1):173–185, 2024.

[Lucey *et al.*, 2010] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *CVPR Workshops*, pages 94–101, 2010.

[Ma *et al.*, 2024] Chengcheng Ma, Ismail Elezi, Jiankang Deng, Weiming Dong, and Changsheng Xu. Three heads are better than one: Complementary experts for long-tailed semi-supervised learning. In *AAAI*, pages 14229–14237, 2024.

[Menon *et al.*, 2021] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, pages 1–24, 2021.

[Mollahosseini *et al.*, 2017] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[Oh *et al.*, 2022] Youngtaek Oh, Dong-Jin Kim, and In So Kweon. Daso: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning. In *CVPR*, pages 9786–9796, 2022.

[Papyan *et al.*, 2020] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

[Roy and Etemad, 2024] Shuvendu Roy and Ali Etemad. Exploring the boundaries of semi-supervised facial expression recognition using in-distribution, out-of-distribution, and unconstrained data. *IEEE Transactions on Affective Computing*, 2024.

[Ruan *et al.*, 2021] Delian Ruan, Yan Yan, Shenqi Lai, Zhenhua Chai, Chunhua Shen, and Hanzi Wang. Feature decomposition and reconstruction learning for effective facial expression recognition. In *CVPR*, pages 7660–7669, 2021.

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

[Tarvainen and Valpola, 2017] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NIPS*, pages 1195–1204, 2017.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11):2579–2605, 2008.

[Wang *et al.*, 2020] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *CVPR*, pages 6897–6906, 2020.

[Wei and Gan, 2023] Tong Wei and Kai Gan. Towards realistic long-tailed semi-supervised learning: Consistency is all you need. In *CVPR*, pages 3469–3478, 2023.

[Wei *et al.*, 2021] Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *CVPR*, pages 10857–10866, 2021.

[Wen *et al.*, 2016] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515, 2016.

[Wu and Cui, 2023] Zhiyu Wu and Jinshi Cui. La-net: Landmark-aware learning for reliable facial expression recognition under label noise. In *ICCV*, pages 20698–20707, 2023.

[Xue *et al.*, 2021] Fanglei Xue, Qiangchang Wang, and Guodong Guo. Transfer: Learning relation-aware facial expression representations with transformers. In *ICCV*, pages 3601–3610, 2021.

[Yang *et al.*, 2022] Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? In *NeurIPS*, pages 37991–38002, 2022.

[Yu *et al.*, 2023] Zhuoran Yu, Yin Li, and Yong Jae Lee. InPL: Pseudo-labeling the inliers first for imbalanced semi-supervised learning. In *ICLR*, pages 1–17, 2023.

[Zeng *et al.*, 2022] Dan Zeng, Zhiyuan Lin, Xiao Yan, Yuting Liu, Fei Wang, and Bo Tang. Face2exp: Combating data biases for facial expression recognition. In *CVPR*, pages 20291–20300, 2022.

[Zhang *et al.*, 2016] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[Zhang *et al.*, 2021] Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. In *NeurIPS*, pages 17616–17627, 2021.

[Zhang *et al.*, 2023] Yuhang Zhang, Yaqi Li, Lixiong Qin, Xuannan Liu, and Weihong Deng. Leave no stone unturned: Mine extra knowledge for imbalanced facial expression recognition. In *NeurIPS*, pages 14414–14426, 2023.