

Towards Semi-Supervised Deep Facial Expression Recognition with An Adaptive Confidence Margin

Hangyu Li¹, Nannan Wang^{1*}, Xi Yang¹, Xiaoyu Wang², Xinbo Gao³

¹ State Key Laboratory of Integrated Services Networks, Xidian University

² The Chinese University of Hong Kong (Shenzhen)

³ Chongqing University of Posts and Telecommunications

hangyuli.xidian@gmail.com, nnwang@xidian.edu.cn, yangx@xidian.edu.cn

fanguhuaxue@gmail.com, gaoxb@cqupt.edu.cn

Abstract

Only parts of unlabeled data are selected to train models for most semi-supervised learning methods, whose confidence scores are usually higher than the pre-defined threshold (i.e., the confidence margin). We argue that the recognition performance should be further improved by making full use of all unlabeled data. In this paper, we learn an **Adaptive Confidence Margin (Ada-CM)** to fully leverage all unlabeled data for semi-supervised deep facial expression recognition. All unlabeled samples are partitioned into two subsets by comparing their confidence scores with the adaptively learned confidence margin at each training epoch: (1) subset I including samples whose confidence scores are no lower than the margin; (2) subset II including samples whose confidence scores are lower than the margin. For samples in subset I, we constrain their predictions to match pseudo labels. Meanwhile, samples in subset II participate in the feature-level contrastive objective to learn effective facial expression features. We extensively evaluate Ada-CM on four challenging datasets, showing that our method achieves state-of-the-art performance, especially surpassing fully-supervised baselines in a semi-supervised manner. Ablation study further proves the effectiveness of our method. The code will be publicly available.

1. Introduction

Facial expression recognition (FER) aims to make computers understand visual emotion. Recently, the advancement of deep FER is largely promoted by large-scale labeled datasets, e.g., RAF-DB [15] and AffectNet [22]. However, the collection of large-scale labels is quite expensive and difficult. Besides, existing labels often fail to

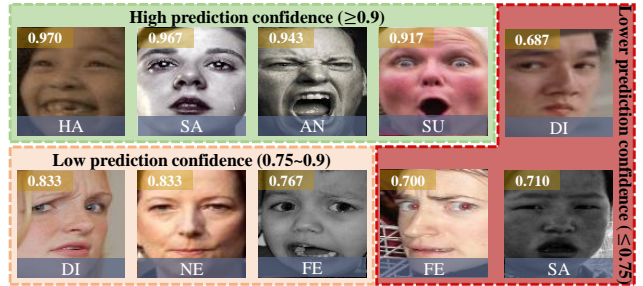


Figure 1. Confidence scores by 30 volunteers on ten faces annotated with seven classes, including surprise, fear, disgust, happiness, sadness, anger and neutral. The upper left corner of each face is tagged with its confidence score. All faces are divided into three groups based on the confidence score. The results provide insights that the confidence scores may be inconsistent among different classes and even the confidence gap between intra-class expressions may be large, e.g., faces annotated with Sadness.

satisfy actual fine-grained needs and the re-labeling data requires human experts. Therefore, it is urgent to develop a powerful method for training models on a large amount of data without corresponding labels, i.e., semi-supervised deep facial expression recognition (SS-DFER).

Most recent semi-supervised learning (SSL) algorithms achieve competitive performance by predicting artificial labels of unlabeled data. For example, pseudo-labeling methods [12, 14, 24, 34] utilize the model predictions as artificial labels to retrain CNN models. Typically, FixMatch [28] explores weakly-augmented and strongly-augmented data pairs and selects only unlabeled samples with high-confidence predictions, whose confidence scores are above the pre-defined fixed threshold (e.g., 0.95).

Despite excellent performance on common classification tasks, the threshold-based pseudo-labeling strategy is still challenging for SS-DFER mainly due to two reasons: (1)

*Corresponding author

The fixed threshold for all categories. Facial expressions from different categories are classified with varying degrees of difficulty. To better understand this, we randomly pick several images from RAF-DB [15] and conduct a user study. As shown in Figure 1, for the face annotated with *Happiness*, the confidence score is much higher than other facial expressions. Especially, the confidence gap between the most and least possibles is up to 28%. Therefore, the fixed threshold is unfair to different facial expressions. In other words, the fixed threshold (*e.g.*, 0.95) may lead to selecting too many expressions with high confidence scores (*e.g.*, happiness) and too few expressions with low or lower confidence scores (*e.g.*, disgust). Moreover, the fixed setting is not adaptive enough at each training epoch. (2) Inefficient data utilization. There is a large gap between the confidence scores of different intra-class samples. For example, the confidence gap between faces annotated with *Sadness* is as large as 25% (see Figure 1). This issue may cause that some intra-class samples with low confidence scores cannot be selected for training models, *e.g.*, the *Sadness* with the confidence score of 0.71. This inspires us to consider that how samples with low confidence scores contribute to feature learning. Hence, **to fully leverage unlabeled data with the adaptive threshold is crucial for SS-DFER.**

To this end, we propose a semi-supervised DFER algorithm with an **Adaptive Confidence Margin** (Ada-CM) to enjoy its adaptivity in terms of the learning on all unlabeled data. Specifically, the proposed Ada-CM firstly runs over all given labeled data and adaptively updates the confidence margin based on the learning difficulty of different facial expressions. Importantly, the confidence margin is gradually improved over training epochs. Then, it predicts confidence scores of weakly-augmented unlabeled data, which are compared with the learned confidence margin to partition all unlabeled samples into two subsets: subset I including samples with high confidence scores (*i.e.*, whose confidence scores are not lower than the margin) and subset II including samples with low confidence scores (*i.e.*, whose confidence scores are lower than the margin). For samples in subset I, Ada-CM leverages strongly-augmented unlabeled samples and pseudo labels from their weakly-augmented versions to calculate the cross-entropy loss. Moreover, for subset II, we conduct a feature-level contrastive objective to learn effective features by applying the InfoNCE loss [4]. Overall, our main contributions can be summarized as follows:

- We propose a novel end-to-end semi-supervised DFER method by adaptively learning the confidence margin. To the best of our knowledge, this is the first solution to explore the dynamic confidence margin in SS-DFER.
- An adaptive confidence margin is designed to dynamically learn on all unlabeled data for the model’s training. More importantly, samples with low confidence

scores are leveraged to enhance the feature-level similarity.

- Extensive experiments on four challenging datasets show the effectiveness of our proposed Ada-CM. Especially, our method achieves superior performance, surpassing fully-supervised baselines in a semi-supervised manner.

2. Related Work

2.1. Facial Expression Recognition

Numerous FER methods [15, 20, 27, 35] have been proposed. There are two major lines of research on FER, *i.e.*, handcraft features and deep learning-based DFER methods.

Traditionally, early attempts [11, 20, 23] focus on the texture information on in-the-lab FER datasets, *e.g.*, CK+ [19] and Oulu-CASIA [41]. Motivated by large-scale unconstrained FER datasets [1, 15, 22], DFER algorithms design effective CNN networks or loss functions to achieve superior performance. Right from the beginning, Li *et al.* [15] proposed a locality preserving loss to learn more discriminative facial expression features. Inspired by the attention mechanism, Wang *et al.* [31] proposed region-based attention network to capture important facial regions. Li *et al.* [18] explored partially-occluded facial expression recognition. Moreover, several works [27, 30, 38] considered the inconsistent annotation problem in DFER. Besides, Xue *et al.* [35] first explored relation-aware representations for Transformers-based DFER.

The above methods perform FER in a fully-supervised manner. Differently, Florea *et al.* [7] proposed an extension of MixMatch [3], namely Margin-Mix, and leveraged unlabeled samples to solve the dense area problem. Indeed, Margin-Mix determined artificial labels of unlabeled samples by the embeddings for class centers, not by the confidence margin. Moreover, the center updating is costly and time-consuming. To the best of our knowledge, no threshold-based pseudo-labeling method has been proposed for the SS-DFER task. In our work, an adaptive confidence margin is designed to produce high-quality pseudo labels of unlabeled samples with high confidence scores.

2.2. Semi-Supervised Learning

In recent years, semi-supervised learning methods have been successfully applied to some challenging problems [28, 32, 39]. Existing works on SSL deploy consistency regularization [26, 33], entropy minimization [8, 14] and traditional regularization [3] to leverage unlabeled data.

Among them, pseudo-labeling is a pioneer SSL method to obtain hard labels from model predictions. Especially, the threshold-based methods [25, 28] select unlabeled samples with high-confidence predictions. FixMatch [28] and UDA [33] obtained pseudo labels based on the fixed

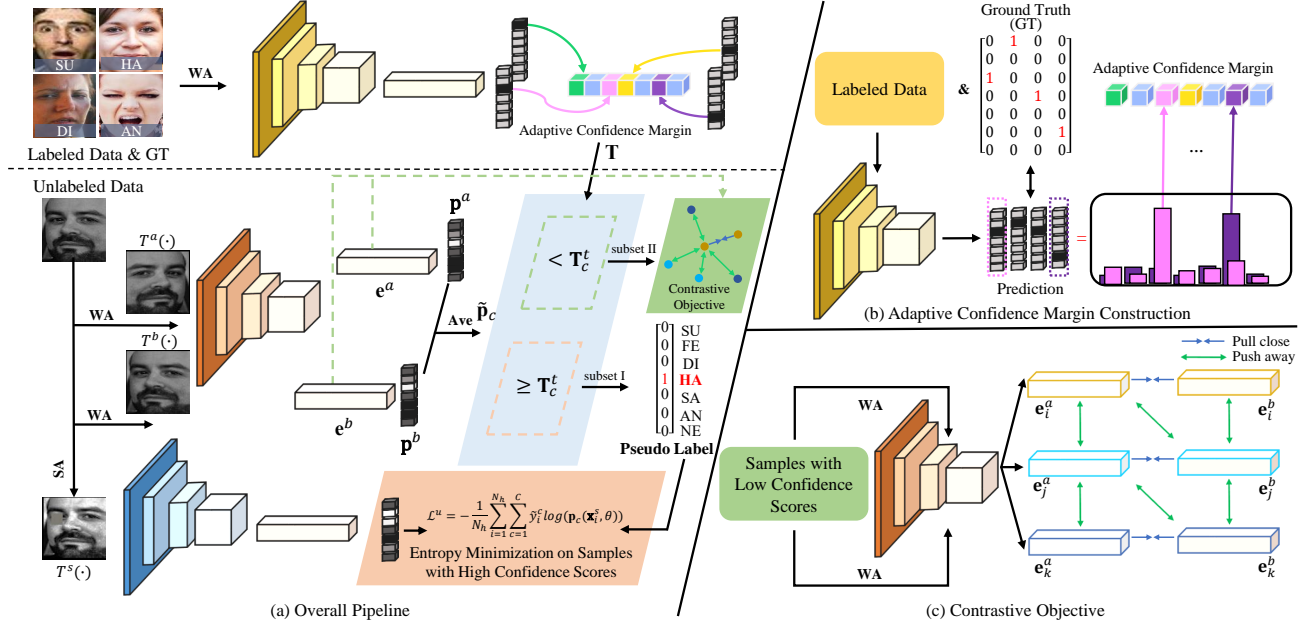


Figure 2. Illustration of Ada-CM. In each forward pass, weakly-augmented (WA) labeled samples are fed into the model to learn the adaptive confidence margin. Specifically, when the model’s prediction is equal to the ground truth, the corresponding confidence scores are put into the confidence margin and then the average is used as the learned margin. Next, two WA unlabeled samples are fed separately into the model, resulting in probability distributions \mathbf{p}^a and \mathbf{p}^b . Then, Ada-CM partitions all unlabeled data into two subsets based on the relationship between the confidence score (*i.e.*, the maximum value in the average probability distribution $\tilde{\mathbf{p}}_c$) and the confidence margin \mathbf{T}_c^t . Finally, samples in subset I with pseudo labels and the feature similarity on samples in subset II are explored by entropy minimization and contrastive objective, respectively. For clarity, we present the same model with three colors to distinguish different inputs.

threshold and leveraged weak and strong augmentations to achieve the consistency regularization. In addition, several works have investigated on the dynamic threshold [34, 39]. For example, Xu *et al.* [34] proposed a generic method to dynamically select samples with high-confidence predictions. In our work, it is the first attempt to learn an adaptive confidence margin for SS-DFER. Besides, all unlabeled samples are learned, which is also the first attempt for SSL.

3. Method

3.1. Problem Formulation

Generally, for a C -class fully-supervised DFER task, there is a set of instance-label pairs as $\mathcal{C} = (\mathcal{X}, \mathcal{Y})$, where $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ and $\mathcal{Y} = \{\mathbf{y}_i \in \{0, 1\}^C\}_{i=1}^N$ are the set of training data and the corresponding one-hot labels, and N denotes the number of labeled training data. The conventional loss function is the cross-entropy loss on the labeled training data:

$$\mathcal{L}_{CE}^s = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C \mathbf{y}_i^c \log(\mathbf{p}_c(\mathbf{x}_i, \theta)), \quad (1)$$

where $\mathbf{p}_c(\mathbf{x}_i, \theta)$ denotes the prediction probability of data \mathbf{x}_i for class c with the model parameter θ . However, for the

problem of semi-supervised DFER, labels are not guaranteed to be fully available. In general, the original training samples are partitioned into two sets, including a labeled set and an unlabeled set. Let

$$\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \dots, N_s\} \quad (2)$$

be the labeled training set. N_s is the number of labeled training data. Besides the labeled set \mathcal{S} , the unlabeled training set shares the same categories, denoted by

$$\mathcal{U} = \{\mathbf{x}_i^u, i = 1, \dots, N_u\}, \quad (3)$$

where N_u is the number of unlabeled training data.

Given the above data, existing pseudo-labeling methods [14, 28, 33] aim to generate the pseudo label $\tilde{\mathbf{y}}_i$ for an unlabeled sample \mathbf{x}_i^u . Then, the model is optimized on the labeled set \mathcal{S} and the unlabeled set \mathcal{U} with pseudo labels by the cross-entropy loss. For example, FixMatch [28] adopts a fixed threshold for all categories and selects samples with high-confidence predictions whose confidence scores are above the threshold. Crucially, for the consistency regularization [3] in SSL, FixMatch conducts two separate weakly-augmented (WA) and strongly-augmented (SA) operations and estimates pseudo labels based on the WA data.¹

¹It is a form of consistency regularization in which the model should output the same prediction for the WA and SA data.

More importantly, the quality of pseudo labels depends on the threshold, which can determine the level of confidence scores. However, existing methods can only make sure that samples with high confidence scores are used for the model’s training. In addition, many expressions (*e.g.*, happiness) usually have higher confidence scores than other expressions, which is unfair to other categories. In this work, we focus on the confidence margin-based pipeline and leverage all unlabeled data regardless of the degree of confidence scores.

3.2. Our proposed Ada-CM

In this section, we first present the overall framework in Sec. 3.2.1. In Sec. 3.2.2, we propose an adaptive confidence margin, which contains different thresholds for expression categories. Furthermore, we introduce the learning on all unlabeled data in Sec. 3.2.3. Finally, we display the whole training objective in Sec. 3.2.4.

3.2.1 The Overall Framework

To fully leverage unlabeled data, we propose a semi-supervised DFER method (see Figure 2). Unlike the fixed threshold for all categories, we propose an adaptive confidence margin (Ada-CM), which consists of different thresholds for each expression category. Then, our proposed Ada-CM partitions all unlabeled data into two subsets by comparing the confidence scores² with the margin. Once the confidence score of unlabeled data (*i.e.*, the maximum value in the average probability distribution $\tilde{\mathbf{p}}_c$) is higher than the corresponding threshold in the margin, the prediction on the SA version will match the pseudo label from the above WA versions via the cross-entropy loss. Otherwise, the contrastive objective is used to enhance the similarity between two WA features. Therefore, our Ada-CM mainly contains two components, including learning an adaptive confidence margin and adaptively learning on unlabeled data. We will elaborate on key technologies in turn.

3.2.2 Adaptive Confidence Margin

Recent SSL progresses [14, 28, 33] select samples with high confidence scores to update models based on a pre-defined threshold for all categories. However, since the confidence score varies by category, it is unfair to different facial expressions. Motivated by this, we aim to evaluate the confidence margin based on given labeled data and build an adaptive confidence margin. Note that our method requires no extra labeled data to determine the margin.

²For labeled and unlabeled data, the confidence score can be viewed as the probability value corresponding to the ground truth and the maximum value in a probability distribution, respectively.

For the labeled set $\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \dots, N_s\}$, we would like to explore the confidence margin for different facial expressions. A classical idea is to calculate the predictions of all labeled samples and obtain different thresholds by averaging the corresponding confidence scores. However, this practice shows a fatal problem for SS-DFER. In particular, several studies have shown that noisy labels exist in DFER datasets [27, 30], which indicates that certain confidence scores of samples are not desirable. Therefore, we propose the adaptive confidence margin based on the *correct* confidence scores.

Specifically, we first calculate predictions of labeled samples and determine predicted labels. Compared to the ground truth $\{\mathbf{y}_i \in \{0, 1\}^C, i = 1, 2, \dots, N_s\}$, we pick out the correctly-predicted samples $\mathcal{S}^T = \{(\mathbf{x}_i, \hat{\mathbf{y}}_i, s_i), i = 1, \dots, N_{st}\}$, where s_i is the confidence score of the i -th labeled data, $\hat{\mathbf{y}}_i \in \{1, 2, \dots, C\}$ denotes the i -th label and N_{st} is the number of data in \mathcal{S}^T . We then construct the adaptive confidence margin $\mathbf{T} = \{(\mathbf{T}_1, \dots, \mathbf{T}_C) | \mathbf{T}_c \in \mathbb{R}, c = 1, \dots, C\}$ by

$$\mathbf{T}_c = \frac{1}{N_{st}^c} \sum_{i=1}^{N_{st}} \mathbb{1}(\hat{\mathbf{y}}_i = c) \cdot s_i, \quad (4)$$

where N_{st}^c reflects the number of samples annotated with the c -th class in \mathcal{S}^T . It is known that with the increasing of epoch, the discriminative ability of DFER model is stronger. Therefore, we consider that the confidence margin is also adaptively improved with the training epoch. Mathematically, the current confidence margin at epoch t is given by

$$\mathbf{T}_c^t = \frac{B\mathbf{T}_c}{1 + \gamma^{-t}}, \quad (5)$$

where $0 < B < 1$ and $\gamma > 1$ are two constants. In practice, we set $B = 0.97$ to control too large margin. Moreover, we use $\gamma = e$ as the default setting. The ablation study about B and γ will be shown in Sec. 4.2.

3.2.3 Adaptively Learning on Unlabeled Data

The proposed adaptive confidence margin is an important criterion to determine the level of confidence scores. To leverage all unlabeled samples efficiently, we design an adaptive learning strategy to explore all unlabeled data for updating model parameters.

To this end, we propose to adaptively learn on all unlabeled data based on the above adaptive confidence margin. Specifically, we first generate two WA versions $\mathbf{x}_i^a = T^a(\mathbf{x}_i^u)$ and $\mathbf{x}_i^b = T^b(\mathbf{x}_i^u)$ and use the model to extract facial expression features and probability distributions. Based on two probability distributions \mathbf{p}^a and \mathbf{p}^b , we then compute the average probability distribution:

$$\tilde{\mathbf{p}}_c = \frac{1}{2}(\mathbf{p}^a(\mathbf{x}_i^a, \theta) + \mathbf{p}^b(\mathbf{x}_i^b, \theta)), \quad (6)$$

where $\tilde{\mathbf{p}}_c$ denotes the probability distribution of data \mathbf{x}_i^u about class c . Now, the adaptive learning strategy compares two values, *i.e.*, $\max(\tilde{\mathbf{p}}_c)$ and $\mathbf{T}_{argmax_c \tilde{\mathbf{p}}_c}^t$, to dynamically partition all unlabeled data into the subset I including samples with high confidence scores and the subset II including samples with low confidence scores.

For samples in subset I, we retain the average as the pseudo label unaltered at the current epoch, *i.e.*, $\tilde{\mathbf{y}}_i = argmax_c \tilde{\mathbf{p}}_c$, where $\tilde{\mathbf{y}}_i$ denotes the one-hot label for convenience. To achieve the consistency regularization, we adopt the strongly-augmented operations and make the prediction of SA version match the pseudo label obtained from two WA versions. Therefore, given a high-confidence sample \mathbf{x}_i^u , the unsupervised loss \mathcal{L}^u is defined as the cross-entropy loss between the SA version $\mathbf{x}_i^s = T^s(\mathbf{x}_i^u)$ and $\tilde{\mathbf{y}}_i$:

$$\mathcal{L}^u = -\frac{1}{N_h} \sum_{i=1}^{N_h} \sum_{c=1}^C \tilde{\mathbf{y}}_i^c \log(\mathbf{p}_c(\mathbf{x}_i^s, \theta)), \quad (7)$$

where N_h denotes the number of data in subset I.

For samples in subset II, since the low-confidence predictions are not believable, the cross-entropy loss cannot be used to guide the model's learning. Inspired by contrastive learning [4, 16, 36], we consider the relationship between two WA versions of the same unlabeled data to improve the discriminative power of expression features. Specifically, the feature-level similarity is first measured by

$$s(\mathbf{e}_i^a, \mathbf{e}_i^b) = \frac{(\mathbf{e}_i^a)(\mathbf{e}_i^b)^\top}{\|\mathbf{e}_i^a\| \|\mathbf{e}_i^b\|}, \quad (8)$$

where \mathbf{e}_i^a and \mathbf{e}_i^b are two weak-augmented expression features. Based on the obtained similarity measure, the contrastive objective for the feature \mathbf{e}_i^a of a sample \mathbf{x}_i^u can be defined as follows:

$$\mathcal{L}^c = -\frac{1}{N_l} \sum_{i=1}^{N_l} \log \left(\frac{e^{s(\mathbf{e}_i^a, \mathbf{e}_i^b)/\tau}}{\sum_j e^{s(\mathbf{e}_i^a, \mathbf{e}_j^a)/\tau} + \sum_k e^{s(\mathbf{e}_i^a, \mathbf{e}_k^b)/\tau}} \right), \quad (9)$$

where $i, k \in I = \{1, 2, 3, \dots, N_l\}$, $j \in I \setminus \{i\}$, N_l is the number of data in subset II and τ is a temperature parameter to control the softness [10]. Notably, this process can further increase the discriminative power of expression features and introduce no additional trainable parameters.

3.2.4 Overall Objective Function

The proposed SS-DFER method with the adaptive confidence margin can be optimized in an end-to-end process. The whole network minimizes the following loss function:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{CE}^s + \lambda_2 \mathcal{L}^u + \lambda_3 \mathcal{L}^c, \quad (10)$$

Algorithm 1 Ada-CM's main learning algorithm.

Input: Model parameters θ , labeled samples and their labels $\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \dots, N_s\}$, unlabeled samples $\mathcal{U} = \{\mathbf{x}_i^u, i = 1, \dots, N_u\}$, number of epoch t_{max} and learning rate η .

Output: Updated model parameters θ .

```

1: // Learning the adaptive confidence margin.
2: Initialization:  $\mathbf{T}^0 \in \{f\}^C$ .
3: for  $i = 1, 2, 3, \dots, N_s$  do
4:   Obtain the correctly-predicted set  $\mathcal{S}^T$ .
5:   Update  $\mathbf{T}_c$  by Eq. (4).
6: end for
7: Obtain the current confidence margin  $\mathbf{T}_c^t$  by Eq. (5).
8: // Training models with labeled and unlabeled samples.
9: for  $t = 1, 2, 3, \dots, t_{max}$  do
10:  Compute  $\mathcal{L}_{CE}^s$  using labeled samples by Eq. (1).
11:  Predict  $\mathbf{p}^a, \mathbf{p}^b$  and the average  $\tilde{\mathbf{p}}_c$  by Eq. (6).
12:  if  $\max(\tilde{\mathbf{p}}_c) \geq \mathbf{T}_{argmax_c \tilde{\mathbf{p}}_c}^t$  then
13:    Compute  $\mathcal{L}^u$  using subset I by Eq. (7).
14:    Update  $\theta \leftarrow \theta - \eta \nabla \mathcal{L}^u$ .
15:  else
16:    Compute  $\mathcal{L}^c$  using subset II by Eqs. (8) and (9).
17:    Update  $\theta \leftarrow \theta - \eta \nabla \mathcal{L}^c$ .
18:  end if
19:  Update  $\theta \leftarrow \theta - \eta \nabla \mathcal{L}_{CE}^s$ .
20: end for
```

where \mathcal{L}_{CE}^s and \mathcal{L}^u denote the cross-entropy loss on labeled samples and unlabeled samples in subset I, respectively. \mathcal{L}^c denotes the contrastive objective on samples in subset II. λ_1 , λ_2 and λ_3 are hyper-parameters to balance each term's intensity. The whole process of our proposed method is summarized in Algorithm 1.

3.3. Discussion

Here, we discuss the relations between the proposed Ada-CM, FixMatch [28], Dash [34] and FlexMatch [39], which share similar philosophy but with different roles.

Relation to FixMatch [28]. FixMatch focuses on the pre-defined fixed threshold so its modeling capacity is limited at the early training stage [34, 39]. Ada-CM aims at the adaptive confidence margin, which is friendly for the early training. In addition, FixMatch selects unlabeled samples with high-confidence predictions through the same threshold for all categories, while Ada-CM can adaptively leverage all unlabeled data based on the learning difficulty of different facial expressions.

Relation to Dash [34]. Dash devotes to selecting the unlabeled samples whose loss values are smaller than the dynamic threshold. However, Ada-CM adopts the confidence score as the compared variable, which intuitively reflects the predictions of unlabeled samples. Furthermore, our adaptive confidence margin is built on different-class labeled samples, while Dash applies the entire labeled set to obtain a unique threshold value.

Relation to FlexMatch [39]. FlexMatch first considers the learning difficulties of each class and selects more unlabeled samples with high-confidence predictions at the early stage. While in Ada-CM, our method can explore all unlabeled data at any learning stage, which significantly improves the efficiency of data utilization.

4. Experiments

In this section, extensive experiments are conducted to verify the effectiveness of our proposed method. We first briefly introduce the experiment setup (Sec. 4.1). Then, we perform the ablation study (Sec. 4.2) to show the importance of each component in Ada-CM. Finally, we compare our method with state-of-the-art methods (Secs. 4.3 to 4.5).

4.1. Experiment Setup

Datasets. We evaluate Ada-CM on four commonly used datasets: RAF-DB, SFEW, AffectNet and CK+. **RAF-DB** [15] includes nearly 30,000 facial images with two different subsets by 40 annotators. In our experiments, we choose the single-label subset with six basic expressions (*i.e.*, surprise, fear, disgust, happy, sad and anger) and the neutral face, which is divided into the training set and testing set with the size of 12,271 and 3,068, respectively. **SFEW** [6] is a static facial expression dataset selected from movies, including 958 images for training, 436 images for validation and 372 images for testing. The images in SFEW are annotated with six basic expressions and the neutral face as in RAF-DB. For the reason that there are no public labels in the testing set, we compare performance on the validation set. **AffectNet** [22] is currently the largest real-world facial expression dataset, consisting of about 420,000 manually-annotated images with eight expression labels. For a fair comparison, we utilize 280,000 training images and 4,000 validation images (500 images per class). **The Extended Cohn-Kanade (CK+)** [19] includes 593 video sequences from 123 subjects. We select the first and last frame of each sequence as the neutral face and targeted expression, including 636 images with seven expression labels.

Performance Metrics. For evaluating the model performance, we utilize the overall test accuracy as the performance metric for all algorithms. Besides, we follow the standard SSL evaluation protocol and perform experiments five times using different random seeds to obtain the mean accuracy and their standard deviations.

Implementation Details. In the following experiments, we use MTCNN [40] to detect and resize facial expressions with the size 224×224 . Our proposed method is implemented with the PyTorch toolbox on two NVIDIA Tesla V100 GPUs. For the backbone CNN, we use the ResNet-18 [9] pre-trained on MS-Celeb-1M face recognition dataset by default. We also conduct experiments with WideResNet-28-2 [37] used in MarginMix [7] for a fair comparison. We

Table 1. Ablation study of the fixed threshold and different components in Ada-CM on RAF-DB and SFEW (in %, mean \pm standard deviation). Baseline denotes that the model is only trained by \mathcal{L}_{CE}^s with limited labeled data. This also applies to the following tables. Note that \mathcal{L}^u denotes different thresholds for obtaining data with high confidence scores, *e.g.*, fixed (rows 2 to 4), dynamic (row 5) and our adaptive confidence margin (rows 6 and 8).

Method	\mathcal{L}^u	\mathcal{L}^c	RAF-DB	SFEW
			100 labels	400 labels
Baseline	-	-	52.43 \pm 2.24	43.85 \pm 2.83
FT = 0.5	✓	-	57.49 \pm 1.77	47.85 \pm 1.89
FT = 0.8	✓	-	58.94 \pm 2.05	48.58 \pm 1.32
FT = 0.95	✓	-	60.67 \pm 2.25	50.37 \pm 0.45
FlexMatch [39]	✓	-	61.23 \pm 2.27	50.99 \pm 1.45
Ada-CM	✓	-	61.50 \pm 2.10	51.04 \pm 0.58
	-	✓	54.27 \pm 2.79	45.99 \pm 0.35
	✓	✓	62.36 \pm 1.10	52.43 \pm 0.67

employ a DFER-related weak augmentation strategy, including *RandomCrop*, *RandomHorizontalFlip* and *Normalize*. Moreover, the RandAugment [5] is used as the strong augmentation scheme following by [28]. The training data in RAF-DB is added in SFEW as additional unlabeled data.

For a fair comparison, we use the Adam optimizer [13] with the initial learning rate of 5×10^{-4} for all experiments. The total number of training epochs is set to 20. The mini-batch size of labeled data is 16 except for AffectNet. These setups are the same for all algorithms for fair comparisons. The initial threshold set is empirically set to $\mathbf{T}^0 = \{0.8\}^C$. In the Eq. (10), the hyper-parameters λ_1 , λ_2 and λ_3 are set as 0.5, 1 and 0.1, respectively.

4.2. Ablation Study

In this section, we analyze the contribution of each component in our method. For convenience, we use ‘FT’ to refer to FixMatch [28] with different fixed thresholds in the following experiments.

Effectiveness of each component in Ada-CM. To evaluate the importance of the proposed adaptive confidence margin, we carry out the ablation study to investigate the \mathcal{L}^u with samples with high confidence scores and \mathcal{L}^c with samples with low confidence scores on RAF-DB with 100 labels and SFEW with 400 labels. In addition, the relation in Section 3.3 can also be verified.

As shown in Table 1, several observations can be summarized as follows. Firstly, compared with the baseline, other methods (rows 2 to 8) leverage unlabeled samples and significantly improve the baseline performance on two evaluation schemes. In all cases, our final Ada-CM (row 8) achieves the best performance improvement. Moreover, different fixed thresholds affect the quality of pseudo labels, which is consistent with the effect in FixMatch [28].

Secondly, the effect of the contrastive objective (row 7)

Table 2. Performance comparison with the state-of-the-art SSL methods on RAF-DB, SFEW and AffectNet using ResNet-18 (in %, mean \pm standard deviation). Fully supervised denotes that all labeled training data is used to train the DFER model. This also applies to the following tables. The fully-supervised baseline results are obtained by DLP-CNN [15] on RAF-DB and SFEW, RAN [31] on AffectNet.

Method	RAF-DB				SFEW		AffectNet	
	100 labels	400 labels	2000 labels	4000 labels	100 labels	400 labels	2000 labels	10000 labels
Baseline	52.43 \pm 2.24	67.75 \pm 0.95	78.91 \pm 0.43	81.90 \pm 0.48	33.76 \pm 1.84	43.85 \pm 2.83	47.52 \pm 0.75	53.18 \pm 0.68
Pseudo-Labeling [14]	54.96 \pm 4.24	69.99 \pm 1.81	79.18 \pm 0.27	82.88 \pm 0.49	34.27 \pm 1.67	45.27 \pm 1.32	48.78 \pm 0.67	53.82 \pm 1.29
MixMatch [3]	54.57 \pm 4.16	73.14 \pm 1.40	79.63 \pm 0.91	83.57 \pm 0.49	34.13 \pm 2.58	44.91 \pm 1.87	49.63 \pm 0.49	53.49 \pm 0.47
UDA [33]	58.15 \pm 1.54	72.39 \pm 1.64	81.16 \pm 0.54	83.56 \pm 0.82	39.22 \pm 2.30	48.90 \pm 1.56	50.42 \pm 0.45	56.49 \pm 0.27
ReMixMatch [2]	58.83 \pm 2.34	73.34 \pm 1.82	79.66 \pm 0.66	83.51 \pm 0.18	35.69 \pm 2.73	48.39 \pm 0.71	50.38 \pm 0.63	55.81 \pm 0.34
FixMatch [28]	60.67 \pm 2.25	73.36 \pm 1.59	81.27 \pm 0.27	83.31 \pm 0.33	38.90 \pm 1.90	50.37 \pm 0.45	50.79 \pm 0.37	56.50 \pm 0.43
Ada-CM	62.36\pm1.10	74.44\pm1.53	82.05\pm0.22	84.42\pm0.49	41.88\pm2.12	52.43\pm0.67	51.22\pm0.29	57.42\pm0.43
Fully Supervised	84.13				51.05		52.97	

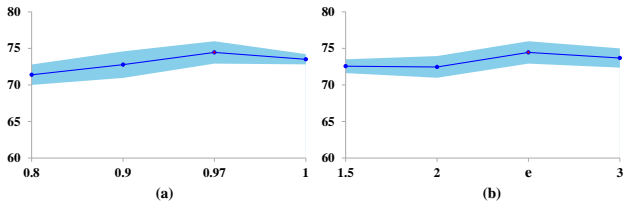


Figure 3. Plots of ablation study on Ada-CM. (a) Varying the control parameter B . (b) Measuring the effect of γ . The performance with default setting is marked in the red.

exceeds the baseline but is not satisfactory. This might be explained by the reason that the contrastive objective focuses on the feature-level similarity between different views of the same data, while having limited ability to distinguish inter-class samples. However, the operation can ensure that all unlabeled samples are leveraged to update models and achieve synergy with \mathcal{L}^u to improve performance.

In addition, for the effect of thresholds, we compare three fixed thresholds, FlexMatch (row 5) and our adaptive confidence margin (row 6). From the results, our adaptive confidence margin is shown to achieve larger performance improvement. These results validate two contributions of our method: 1) Compared with the fixed threshold-based methods, our method is highly effective in pseudo-labeling unlabeled facial expressions. 2) Our Ada-CM and FlexMatch [39] both on samples with high confidence scores achieve similar performance. However, the contribution of our method is that the Ada-CM can leverage all unlabeled samples, compared with selecting only parts of samples in FlexMatch. Indeed, combining the adaptive confidence margin and contrastive objective, our method (row 8) achieves the best results, which demonstrates that with the help of all unlabeled data, entropy minimization and contrastive learning can jointly guide models to extract more discriminative features.

Evaluation of B . Since the parameter B is used to control the peak of confidence margin at each epoch, we con-

Table 3. Performance comparison with the state-of-the-art SS-DFER methods on RAF-DB using WideResNet-28-2 (in %, mean \pm standard deviation).

Method	Labeled samples		
	400	1000	4000
Baseline	26.75	35.25	55.66
MeanTeacher [29]	28.23	36.53	60.36
MixMatch [3]	42.25	60.37	65.24
MarginMix [7]	45.75	66.47	70.68
Ada-CM	59.03\pm0.73	68.38\pm0.44	75.98\pm0.41

duct experiments to explore different B in Eq. (5). Figure 3 (a) reflects the model performance with different B . We find that the default setting $B = 0.97$ achieves the best result. When B is too small, it is difficult for our method to ensure the quality of pseudo labels. The reason is that the amount of data with wrong pseudo labels increases.

Influence of different γ . γ provides the ability to gradually modify the current confidence margin. Figure 3 (b) shows the effects of different $\gamma \in \{1.5, 2, e, 3\}$. We can obtain that our method is not sensitive to γ in a certain range but obtains the top performance when γ is set to e .

4.3. Comparison with State-of-the-Art Methods

To verify the effectiveness of our Ada-CM, we provide experimental results on RAF-DB, SFEW and AffectNet datasets to compare with state-of-the-art methods in two aspects, including comparison with the SS-DFER method [7] on RAF-DB and comparison with SSL methods. Table 2 compares our method with SSL methods using ResNet-18 as the backbone network. From this table, it clearly shows that our proposed Ada-CM achieves the best performance and surpasses the state-of-the-art FixMatch [28] with a large margin. This indicates that our method can better leverage unlabeled data to further improve SSL performance. Compared with the fully-supervised results [15, 31], our method can still beat the baselines with large gains, *i.e.*, 0.29% on RAF-DB, 1.38% on SFEW and 4.45% on AffectNet for the

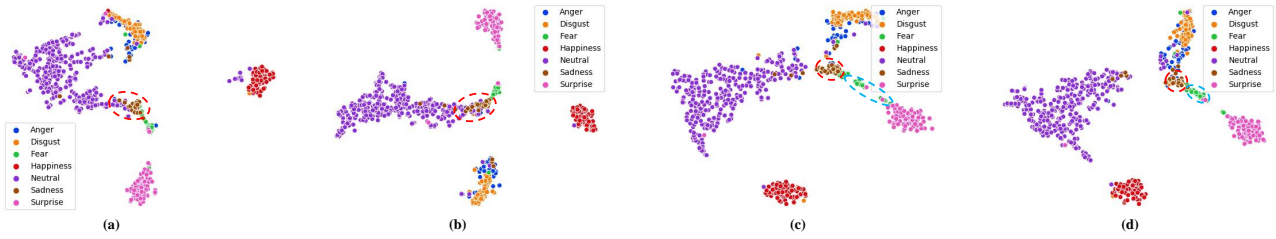


Figure 4. 2D t-SNE visualization [21] of facial expression features obtained by different methods, including (a) Baseline, (b) FixMatch, (c) our Ada-CM (w/o the contrastive objective) and (d) the whole Ada-CM. All models are trained on RAF-DB with 4,000 labels. The features are extracted from the CK+ dataset.

Table 4. Cross-dataset evaluation on in-the-lab CK+ using WideResNet (WRN)-28-2 [37] and ResNet-18 [9] (in %, mean \pm standard deviation). All models are trained on RAF-DB and tested on CK+ dataset.

Method	Labeled samples		Backbone
	100	4000	
Baseline	44.29 \pm 3.04	70.97 \pm 2.21	WRN-28-2
MixMatch [3]	50.42 \pm 8.36	71.76 \pm 1.48	
FixMatch [28]	52.52 \pm 9.69	76.98 \pm 2.15	
Ada-CM	56.13\pm6.85	79.34\pm1.14	
Baseline	59.02 \pm 3.63	80.63 \pm 0.62	ResNet-18
MixMatch [3]	59.94 \pm 5.46	83.87 \pm 1.02	
FixMatch [28]	73.62 \pm 1.78	84.18 \pm 0.99	
Ada-CM	76.92\pm3.57	85.32\pm0.98	
Fully Supervised		81.07 [18]	
		81.72 [17]	

case of 1/3, 1/2 and 1/28 labeled data ratio, respectively. These results verify the effectiveness of our method and the ability to deal with the real-world limited labeled case.

Besides, the proposed Ada-CM can always outperform MarginMix [7] in each case. To the best of our knowledge, MarginMix could be the first attempt to solve the SS-DFER problem based on MixMatch [3]. As shown in Table 3, our Ada-CM significantly surpasses it by 13.27%, 1.91% and 5.3% with 400, 1,000 and 4,000 labeled samples, respectively. The remarkable results demonstrate the effectiveness of our proposed Ada-CM in dealing with SS-DFER. More results could be found in the supplementary material.

4.4. SSL for Cross-Dataset Evaluation

To further verify the generalization ability of our method, we conduct a cross-dataset evaluation scheme (RAF-DB to CK+ dataset), which is widely used in the cross-dataset DFER. Table 4 shows the comparison with state-of-the-art methods using WideResNet-28-2 and ResNet-18 as the backbone. Obviously, our method achieves better performance than existing methods in all cases. Compared with the fully-supervised results [17], the Ada-CM with 4,000 labeled samples using ResNet-18 obtains larger gains by

3.6%. It suggests that our method focuses on a large amount of unlabeled data without the influence of original labels, which is conducive to generalization. Furthermore, our method can achieve superior performance with 1/3 labeled data and fewer model parameters. To be specific, the backbone used in [17] is ResNet-50 with channel level attention, while we use the more lightweight ResNet-18.

4.5. Visualization

To further evaluate the effectiveness of the important adaptive confidence margin in our method, we use t-SNE [21] to visualize the facial expression feature distribution extracted by the baseline, FixMatch, our proposed Ada-CM (w/o the contrastive objective) and the whole Ada-CM on the 2D space, respectively.

As shown in Figure 4, we can observe that the expression features obtained by the baseline and FixMatch are not enough discriminative for some facial expressions, *e.g.*, the sadness in the red dotted line. In contrast, our Ada-CM (w/o the contrastive objective) can achieve a clear boundary between sadness and other categories. Especially, after combining the contrastive objective, the intra-class similarity and inter-class differences are more distinct.

5. Conclusion

In this paper, we propose a novel Adaptive Confidence Margin (Ada-CM) for semi-supervised deep facial expression recognition, which adaptively leverages all unlabeled samples (*i.e.*, samples in subset I with high confidence scores and samples in subset II with low confidence scores) to train models. The proposed Ada-CM dramatically improves the performance from two aspects. On one hand, unlabeled samples whose confidence scores exceed the learned confidence margin are directly pseudo-labeled to match the predictions of strongly-augmented versions. On the other hand, the contrastive objective is applied to learn facial expression features among samples in subset II. Experiments on four popular datasets show the superiority of our method to perform the SS-DFER task.

References

- [1] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 279–283, 2016. [2](#)
- [2] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *International Conference on Learning Representations*, 2020. [7](#)
- [3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. [2](#), [3](#), [7](#), [8](#)
- [4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. [2](#), [5](#)
- [5] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. [6](#)
- [6] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2106–2112, 2011. [6](#)
- [7] Corneliu Florea, Mihai Badea, Laura Florea, Andrei Racoviteanu, and Constantin Vertan. Margin-mix: Semi-supervised learning for face expression recognition. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020. [2](#), [6](#), [7](#), [8](#)
- [8] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*, pages 529–536, 2005. [2](#)
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [6](#), [8](#)
- [10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. [5](#)
- [11] Yuxiao Hu, Zhihong Zeng, Lijun Yin, Xiaozhou Wei, Xi Zhou, and Thomas S Huang. Multi-view facial expression recognition. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–6. IEEE, 2008. [2](#)
- [12] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5070–5079, 2019. [1](#)
- [13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [6](#)
- [14] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on challenges in Representation Learning*, volume 3, page 896, 2013. [1](#), [2](#), [3](#), [4](#), [7](#)
- [15] 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. [1](#), [2](#), [6](#), [7](#)
- [16] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):8547–8555, May 2021. [5](#)
- [17] Yingjian Li, Guangming Lu, Jinxing Li, Zheng Zhang, and David Zhang. Facial expression recognition in the wild using multi-level features and attention mechanisms. *IEEE Transactions on Affective Computing*, pages 1–1, 2020. doi: [10.1109/TAFFC.2020.3031602](#). [8](#)
- [18] 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. [2](#), [8](#)
- [19] 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 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101, 2010. [2](#), [6](#)
- [20] Yuan Luo, Cai-ming Wu, and Yi Zhang. Facial expression recognition based on fusion feature of pca and lbp with svm. *Optik-International Journal for Light and Electron Optics*, 124(17):2767–2770, 2013. [2](#)
- [21] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. [8](#)
- [22] 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. [1](#), [2](#), [6](#)
- [23] Matti Pietikäinen, Abdenour Hadid, Guoying Zhao, and Timo Ahonen. *Computer vision using local binary patterns*, volume 40. Springer Science & Business Media, 2011. [2](#)
- [24] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, 2015. [1](#)
- [25] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. In *IEEE Workshops on Applications of Computer Vision*, volume 1, pages 29–36, 2005. [2](#)
- [26] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in neural information processing systems*, volume 29, pages 1163–1171, 2016. [2](#)
- [27] Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, and Tao Mei. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression

- recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6248–6257, June 2021. 2, 4
- [28] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, pages 596–608, 2020. 1, 2, 3, 4, 5, 6, 7, 8
- [29] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, pages 1195–1204, 2017. 7
- [30] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6897–6906, 2020. 2, 4
- [31] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020. 2, 7
- [32] Zhenyu Wang, Yali Li, Ye Guo, Lu Fang, and Shengjin Wang. Data-uncertainty guided multi-phase learning for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4568–4577, 2021. 2
- [33] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*, volume 33, pages 6256–6268, 2020. 2, 3, 4, 7
- [34] Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *International Conference on Machine Learning*, pages 11525–11536. PMLR, 2021. 1, 3, 5
- [35] Fanglei Xue, Qiangchang Wang, and Guodong Guo. Transfer: Learning relation-aware facial expression representations with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3601–3610, October 2021. 2
- [36] Yazhou Yao, Zeren Sun, Chuanyi Zhang, Fumin Shen, Qi Wu, Jian Zhang, and Zhenmin Tang. Jo-src: A contrastive approach for combating noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5192–5201, 2021. 5
- [37] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 6, 8
- [38] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*, pages 222–237, 2018. 2
- [39] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 2, 3, 5, 6, 7
- [40] 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. 6
- [41] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti Pietikäläinen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011. 2