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

Hangyu Li[1], Nannan Wang[1*], Xi Yang[1], Xiaoyu Wang[2], Xinbo Gao[3]
[1]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
fanghuaxue@gmail.com, gaoxb@cqupt.edu.cn

In this document, we supply some implementation details of our proposed Ada-CM and more experimental results to further verify the effectiveness of our method.

1. Implementation Details

We have described the implementation details for our method in the main text. More details are provided in this section. Firstly, note that the average probability distribution of two weakly-augmented versions is used as the basis for pseudo-labeling. Therefore, for fairness in all experiments, this strategy is applied to other SSL methods, including FixMatch [4] and FlexMatch [7].

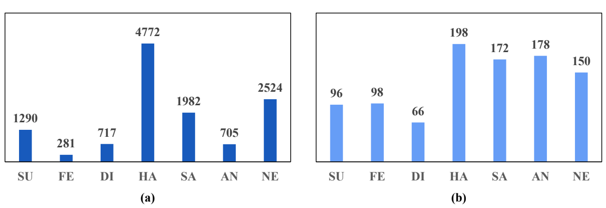

Figure 1. Data distribution of different categories in the training set: (a) RAF-DB and (b) SFEW dataset. (SU=Surprise, FE=Fear, DI=Disgust, HA=Happiness, SA=Sadness, AN=Anger and NE=Neutral)

In addition, most semi-supervised learning (SSL) methods focus on the case of balanced labeled data distribution. However, there is a fact that existing DFER datasets contain some limited facial expressions (*e.g.*, the fear in RAF-DB) making the label distribution highly imbalanced (See Figure 1). Therefore, we will describe data details and list the data distribution of labeled data in our experiments for fair comparisons.

Table 1 shows the data distribution of different-class labeled data, which is applied to RAF-DB and SFEW datasets. For example, for the case of 100 labels, the labeled training set consists of 10 faces annotated with fear and 15 faces annotated with other expressions (*i.e.*, the other six categories). In addition, since AffectNet is the largest dataset, labeled samples are balanced in our experiments.

Table 1. Data distribution of labeled data.

Labels	100	400	1000	2000	4000
Fear	10	40	100	200	250
Others	15	60	150	300	625

2. Ablation Study

Effect of different $\mathbf{T}^0$. $\mathbf{T}^0$ is the initial confidence margin for determining the level of confidence scores at the first epoch. Moreover, considering that the confidence score is not high enough at the early epoch, the initial margin is also used to control the current margin, *i.e.* the margin is no lower than the initial setting. Figure 2 shows the influence of different $\mathbf{T}^0 \in \{0.5, 0.6, 0.75, 0.8, 0.9, 0.95\}^C$.

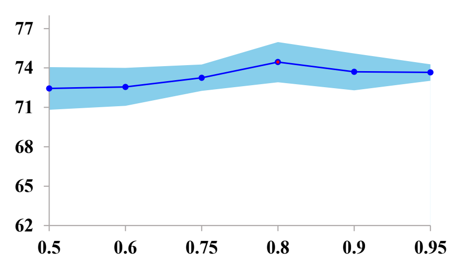

Figure 2. Plots of ablation study on the initial margin $\mathbf{T}^0$. The performance with the default setting is marked in red. The experiments are conducted on RAF-DB with 400 labels, which is the same as the ablation study in the main text.

3. More Comparisons with MarginMix

FERPlus [1] is extended from FER2013, providing a set of new labels created by 10 crowd-sourced annotators. It consists of 28,709 training images, 3,589 validation images and 3,589 testing images. Differently from RAF-DB

*Corresponding author

Table 2. Performance comparison with the state-of-the-art FixMatch [4] and FlexMatch [7] on RAF-DB, SFEW and CK+ using ResNet-18 (in %, mean $\pm$ standard deviation).

Method	RAF-DB		SFEW	CK+	
	400 labels	2000 labels	100 labels	100 labels	4000 labels
Baseline	67.75$\pm$0.95	78.91$\pm$0.43	33.76$\pm$1.84	59.02$\pm$3.63	80.63$\pm$0.62
FixMatch [4]	73.36$\pm$1.59	81.27$\pm$0.27	38.90$\pm$1.90	73.62$\pm$1.78	84.18$\pm$0.99
FlexMatch [7]	73.42$\pm$0.18	81.41$\pm$0.29	40.14$\pm$1.41	75.24$\pm$1.96	84.38$\pm$0.49
Ada-CM	**74.44**$\pm$1.53	**82.05**$\pm$0.22	**41.88**$\pm$2.12	**76.92**$\pm$3.57	**85.32**$\pm$0.98

Table 3. Performance comparison with the state-of-the-art MarginMix [3] on FERPlus using WideResNet-28-2 (in %, mean $\pm$ standard deviation).

Method	Labeled samples		
	320	2000	4000
Baseline	-	50.29	56.78
MeanTeacher [5]	-	50.84	58.28
MixMatch [2]	45.60	58.35	70.91
MarginMix [3]	50.76	60.83	75.18
Ada-CM	**54.61**$\pm$2.17	**73.17**$\pm$1.05	**79.49**$\pm$0.40

and SFEW datasets, FERPlus consists of eight-class facial expressions, including the facial expression of *Contempt*. Since MarginMix [3] conducts experiments on FERPlus, we also supplement the results for a fair comparison.

As shown in Table 3, our Ada-CM also outperforms MarginMix [3] on FERPlus with a large margin, demonstrating that our proposed method can effectively solve the semi-supervised DFER problem.

4. More Comparisons with FlexMatch

In this work, we propose a novel adaptive confidence margin (Ada-CM), which can adaptively leverage all unlabeled facial expressions. To the best of our knowledge, Dash [6] and FlexMatch [7] first investigate the dynamic threshold for SSL. Among them, FlexMatch [7] considers class-related dynamic thresholds, which is closely related to our method. The effectiveness of our proposed adaptive confidence margin has been proved (see the Ablation Study in the main text), *i.e.*, our Ada-CM without the contrastive objective can surpass FlexMatch. Here, we focus on the comparison between FlexMatch and our whole method.

Specifically, we conduct experiments on RAF-DB and SFEW datasets and the cross-dataset evaluation on CK+ with fewer labeled data. Table 2 shows the comparison of the baseline, FixMatch, FlexMatch and our Ada-CM. It is clear that fully leveraging all unlabeled data, our method can achieve better recognition performance.

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 *ACM ICMI*, pages 279–283, 2016.

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, volume 32, pages 5050–5060, 2019.

[3] 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.

[4] 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 *NeurIPS*, volume 33, pages 596–608, 2020.

[5] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, volume 30, pages 1195–1204, 2017.

[6] 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 *ICML*, pages 11525–11536, 2021.

[7] 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 *NeurIPS*, volume 34, 2021.