Exploring Facial Expression Recognition through Semi-Supervised Pre-training and Temporal Modeling

Jun Yu[1], Zhihong Wei[1], Zhongpeng Cai [1]*, Gongpeng Zhao[1], Zerui Zhang[1], Yongqi Wang[1],
Guochen Xie[1], Jichao Zhu[1], Wangyuan Zhu[1] , Qingsong Liu[2], Jiaen Liang[2]
[1] University of Science and Technology of China
[2] Unisound AI Technology Co., Ltd
harryjun@ustc.edu.cn
{weizh588,zpcai,zgp0531,igodrr,wangyongqi,xiegc,
jichaozhu,zhuwangyuan}@mail.ustc.edu.cn
{liuqingsong,liangjiaen}@unisound.com

Abstract

Facial Expression Recognition (FER) plays a crucial role in computer vision and finds extensive applications across various fields. This paper aims to present our approach for the 6th Affective Behavior Analysis in-the-Wild (ABAW) competition, scheduled to be held at CVPR2024. In the facial expression recognition task, the limited size of the FER dataset poses a challenge to the expression recognition model's generalization ability, resulting in subpar recognition performance. To address this problem, we employ a Semi-supervised learning technique to generate expression category pseudo labels for unlabeled face data. At the same time, we uniformly sampled the labeled facial expression samples and implemented a debiased feedback learning strategy to address the problem of category imbalance in the dataset and the possible data bias in semi-supervised learning. Moreover, to further compensate for the limitation and bias of features obtained only from static images, we introduced a Temporal Encoder to capture temporal relationships between neighbouring expression image features. In the 6th ABAW competition, our method achieved the third place in the official test set, a result that fully demonstrates the effectiveness and competitiveness of our proposed method.

1. Introduction

The goal of Facial Expression Recognition (FER) is to identify the emotional state of an individual by analyzing facial images or videos. It is a broad research field spanning multiple fields such as machine learning, image processing, psychology, etc., with a wide range of applications, including safe driving, intelligent monitoring, and human-computer interaction. Given its diverse applications, it is extremely important to establish a robust FER system. The expression recognition task is a classic problem in the field of pattern recognition, which usually involves the classification of six basic expressions: happiness, surprise, sadness, anger, disgust and fear. The facial expression recognition community has made tremendous progress in recent years. State-of-the-art FER methods achieve very good results on numerous public datasets, such as RAF-DB[23], SFEW[37] and AffectNet[26]. To stimulate interdisciplinary collaboration and address pivotal research inquiries spanning affective computing, machine learning, and multi-modal signal processing, Kollias et al. have spearheaded the Affective Behavior Analysis in-the-wild (ABAW) initiative[9–20, 38]. The 6th ABAW workshop and competition is slated to align with the IEEE CVPR conference in 2024.

Traditional fully supervised Facial Expression Recognition (FER) methods rely on the availability of large volumes of high-quality labeled data to fine-tune model precision [1, 2, 32]. However, mainstream training datasets often suffer from class imbalance. Most fully supervised models tend to accurately identify majority classes, thus reducing the model's accuracy for minority classes. Sometimes minority classes constitute less than 10% of the data, and this disparity in data volume makes it challenging for models to learn fairly across all categories. Financial and logistical constraints in acquiring extensive labeled FER data hinder the expansion of training repositories. In contrast, the volume of data for Face Recognition (FR) surpasses that of FER. Making a leap to augment samples from FR data to aid models in learning the FER task could be highly beneficial. Addressing the class imbalance issue and the discrepancy

*Corresponding author

in data distribution between FR and FER data to effectively utilize FR data in support of FER models presents an urgent challenge.

In this study, we propose a two-phase methodology to enhance the recognition and analysis of facial expressions. The first phase, known as the spatial pre-training phase, plays a crucial role in preparing the model for subsequent tasks. This phase is an improvement on the method[39]. During this phase, we leverage the power of semi-supervised learning techniques to generate pseudo labels for expression categories using unlabeled face data. This approach ensures a sufficiently large training corpus, allowing the model to effectively extract robust facial expression features. To tackle the challenges of category imbalance in the dataset and potential data bias during semi-supervised learning, we adopt two strategies. First, we uniformly sample labeled facial expression instances to address the category imbalance. Additionally, we employ a debiased feedback learning strategy to mitigate the impact of potential data bias. These strategies collectively contribute to training a more robust facial expression recognizer. Moving to the second phase, the Temporal Refinement phase, we aim to further improve the recognition and analysis of facial expressions by capturing the temporal dynamics. In this phase, we freeze the facial expression recognizer trained in the first phase, which has already acquired strong spatial representation capabilities. To incorporate temporal information, we introduce a temporal encoder that learns the temporal relationships between neighbouring expression image features. By taking temporal information into account, we not only compensate for the inherent feature bias obtained only from still images but also aggregate information from neighbouring frames to make the overall feature representation of a video frame segment smoother. This integration of temporal dynamics allows us to perform more accurate and comprehensive dynamic recognition and analysis of facial expressions. To sum up, our contributions can be summarized as:

- To address the problem of scarcity of facial expression data, we applied a semi-supervised learning technique to generate expression category pseudo labels for unlabeled face data. At the same time, we uniformly sampled the labeled facial expression samples and implemented a debiased feedback learning strategy to solve the problem of category imbalance in the dataset and the possible data bias in semi-supervised learning.
- To compensate for the limitations and biases of features acquired only from static images, we introduce a temporal encoder to learn and capture temporal relationships between neighbouring expression image features. This strategy aims to enhance the model's ability to recognize and analyze the dynamic changes in facial expressions and achieve more accurate dynamic facial expression recognition.
- In the 6th ABAW competition, our method achieved an excellent third place in the official test set, a result that fully demonstrates the effectiveness and competitiveness of our proposed method.

2. Related Work

2.1. Facial Expression Recognition

The task of recognizing facial expressions is a foundational challenge in pattern recognition. Techniques such as those based on deep learning, attention mechanisms, multitasking, and multimodality leverage fully supervised data have significantly advanced the field of Facial Expression Recognition (FER), as evidenced by groundbreaking research,[2, 23, 30, 34, 44, 45]. However, datasets for facial expressions face notable limitations, including a lack of diversity and significant category imbalances. In response, recent efforts have shifted towards expanding these datasets to enrich the variety of facial expressions available for analysis. One pioneering approach to address the challenge of inconsistent labeling across different facial expression datasets is the IPA2LT framework [40]. This methodology introduced the LTNet scheme, an innovative strategy for uncovering the underlying truths among diverse, inconsistent labels through the use of embedded analysis. In the realm of semi-supervised learning for FER, Ada-CM[22] emerged as the first to investigate the concept of dynamic confidence. By designing an adaptive confidence margin, this approach innovatively adapts during training to maximize learning from unlabeled data through feature-level comparisons, utilizing the InfoNCE loss[36] to capture valuable features effectively. Further advancing the field, Face2Exp[39] introduced the Meta-Face2Exp framework. This methodology utilizes a meta-optimization framework to derive unbiased knowledge from auxiliary Facial Recognition (FR) data. Recently, an innovative technique proposed by Zhang et al.[46] employs rebalancing attention mapping to regularize models. This allows for the extraction of transformation-invariant information from secondary categories across all training samples, addressing the critical issue of data imbalance by focusing on underrepresented categories.

2.2. Semi-Supervised Learning

Semi-supervised learning (SLS) has garnered significant attention from researchers in recent years for its ability to utilize unlabeled data to generate artificial labels and enhance the training of real samples. Consistent regularization, as explored by [21, 28], involves leveraging the predictive power of a model to generate synthetic labels through random modifications of inputs or model processes. A notable approach in this regard is FixMatch[31], which combines the creation of synthetic labels with weak and strong

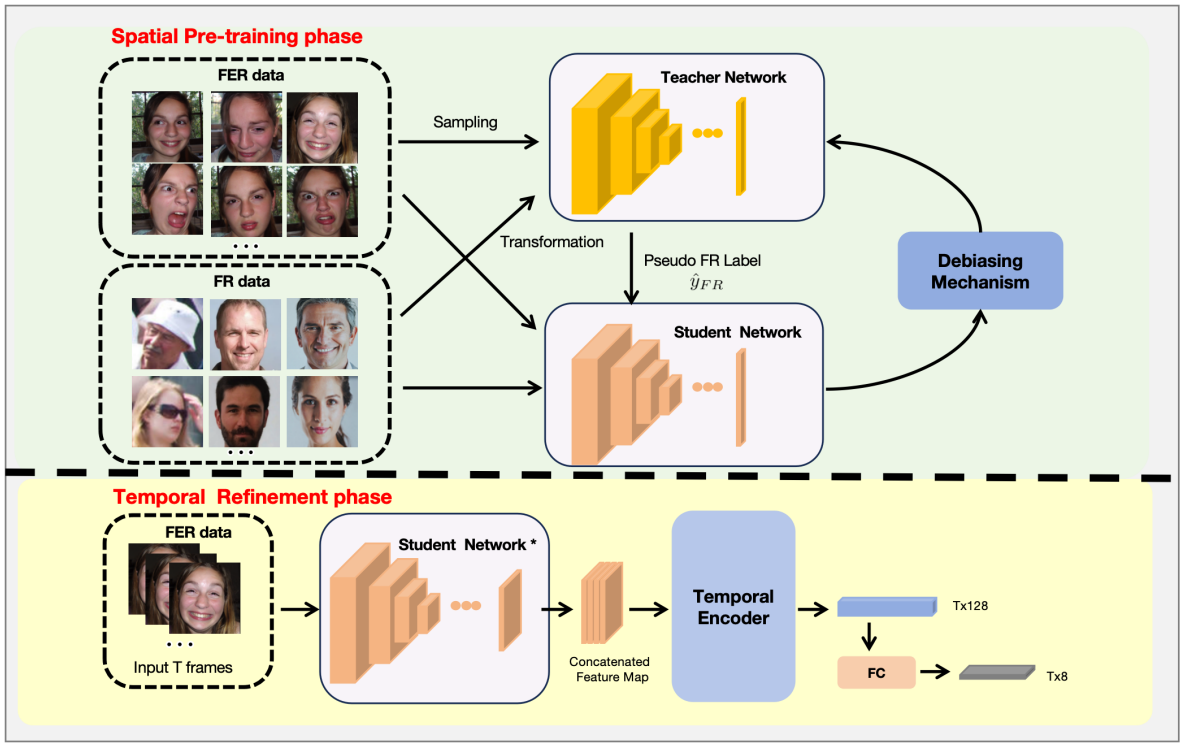

Figure 1. Framework Description. Our approach is mainly divided into a Spatial Pre-training phase and a Temporal Refinement phase. (1). The goal of the Spatial Pretraining phase is to expand face expression data by mining large-scale unlabeled faces through a semi-supervised algorithm. (2). The goal of the Temporal Refine phase is to do temporal feature enhancement of the image features extracted by the student network in the first phase using a temporal encoder, so as to improve the accuracy of recognizing the dynamic facial expressions in video.

data augmentation to achieve consistent regularization and generate pseudo-labeled data based on a certain confidence level. However, FixMatch's effectiveness is hindered during the early training phase due to a fixed threshold. To address this limitation, FlexMatch[41]introduces Curriculum Pseudo Labeling (CPL). CPL adapts to the learning phase of the model and dynamically adjusts the thresholds for different categories, enabling the incorporation of informative unlabeled data and its corresponding pseudo labels. FreeMatch[33] adaptively adjusts the confidence threshold according to the learning state of the model and further introduces adaptive class-fair regularization penalties to encourage the model to make diverse predictions in the early training stage.

3. Method

3.1. Overview

As shown in Figure.1, our approach uses multiple techniques to improve facial expression recognition (FER) in two phases. The first phase, the Spatial Pre-training phase, uses semi-supervised learning to increase the FER dataset by using facial recognition (FR) data. This helps in training a robust image feature extractor. We use two neural networks, a teacher and a student, with the same structure but different weights. In this phase, we improve the networks by using a debiasing method. This involves comparing biased FR data (unlabeled) with debiased FER data (labeled) to create pseudo labels. The teacher network learns from balanced FER data and makes these pseudo labels, which the student network then uses to learn from the FR data, adjusting based on feedback to improve accuracy. In the training, the teacher and student networks update alternately with the debiasing method to better process FR data, improving accuracy and debiasing. The second stage is the temporal refinement stage, where image features are extracted using a student network (frozen after the first stage). To further address the inherent bias introduced by static images, we use a temporal encoder to understand the feature relationships over time, thus improving the accuracy of the model. In addition, the temporal encoder aggregates information from neighboring frames, resulting in a smoother overall feature representation of the video frame segments. This allows for dynamic recognition and analysis of facial expressions in the video, with frame-by-frame predictions provided by the classifier.

3.2. Data Pre-process

3.2.1 Data Sampling

To achieve class balance, we meticulously sample the labeled Facial Expression Recognition (FER) data, ensuring an equal distribution of samples across each expression category. This strategic sampling method allows the model to learn features that are more evenly distributed among classes, significantly contributing to the de-biasing process in Facial Recognition (FR) data analysis.

3.2.2 Data Augmentation

To enhance the diversity and robustness of our dataset, we have employed a data augmentation strategy. In our pre-training process, we have utilized RandAugment[3] as a data augmentation technique to enhance the learning efficiency of our model on training data. By randomly selecting and applying a series of image enhancement operations like rotation, color adjustment, and more, RandAugment has effectively increased the diversity of our data. Initially, we have chosen a milder level of augmentation to ensure smooth training. As the model's performance improves, we gradually increase the intensity of augmentation to further challenge and enhance the model's generalization ability.

3.3. Semi-Supervised Training

The focus of this phase is to utilize semi-supervised learning to extend the FER dataset by using the Facial Recognition (FR) dataset. This expansion helps train an efficient static image feature extractor. In this phase, we employ two neural networks with identical structures but independent weights: the teacher network (T) and the student network (S).

To optimize network performance in this phase, we employ a debiasing mechanism as a core strategy. This involves analyzing the disparities between biased FR data (which is unlabeled) and debiased FER data (which is labeled) to generate pseudo labels. By sampling FER data in a category-balanced manner, the teacher network learns and generates these pseudo labels. Subsequently, the student network employs these pseudo labels to train on the FR data, continuously adjusting itself based on feedback to enhance recognition accuracy. During the training process, the teacher network and the student network are alternately updated, coupled with the de-biasing strategy. This iterative updating aims to improve both the accuracy and de-biasing effect of processing FR data. Consequently, this enhances the prediction ability of the student network.

3.3.1 Student Network

To enhance FER, the student network leverages the rich and comprehensive diversity of large-scale unlabeled FR data. During training, the student network utilizes this unlabeled FR data along with pseudo labels generated by the teacher's network. The training process involves encouraging both networks to predict similar conditional classification distributions for the unlabeled FR data, achieved through the use of the $\mathcal{L}_u$ loss function. The expression for $\mathcal{L}_u$ is given as:

$$\mathcal{L}_u = \mathrm{CE}(\hat{y}_{FR}, \mathcal{S}(x_{FR}; \theta_{\mathrm{s}})). \tag{1}$$

where θ_{s} is the parameter of the student network, and the CE denotes the cross-entropy loss.

3.3.2 Teacher Network

A sampling module, denoted as $\mathrm{Smp}(\cdot)$, is utilized to guarantee a balanced class distribution in the Facial Expression Recognition (FER) dataset. To achieve this, an equal number of samples from each facial expression category were randomly selected, thereby ensuring a balanced class representation for training the teacher network. The learning process integrates three specific types of loss functions: supervised loss, consistency loss, and feedback loss, which collectively provide effective guidance to the teacher network. This process is represented by the following equation:

$$\mathcal{L}_{\mathcal{T}} = \mathcal{L}_s + \mathcal{L}_c + \mathcal{L}_f. \tag{2}$$

For the supervised learning, aimed at minimizing the supervised loss function on a balanced and labeled Facial Expression Recognition (FER) dataset, the supervised loss function $\mathcal{L}_s$ is formulated as follows:

$$\mathcal{L}_s = \mathsf{CE}(y_{FER}, \mathcal{T}(x_{FER}; \theta_{\mathrm{t}})). \tag{3}$$

where θ_{t} is a parameter of the teacher network and CE denotes the cross-entropy loss.

For consistency learning,the teacher network requires that the original image and the augmented counterpart have close class-conditional distributions with a consistency loss function $\mathcal{L}_c$:

$$\mathcal{L}_c = \mathrm{CE}(\mathcal{T}(x_{FR}; \theta_{\mathrm{t}}), \mathcal{T}(\mathrm{Aug}(x_{FR}); \theta_{\mathrm{t}})). \tag{4}$$

The $\mathrm{Aug}(\cdot)$ denotes strong data augmentation, this approach utilizes intensive data augmentation techniques, including rotation, removal, and pixel-level image manipulation.

For the feedback learning, the process involves estimating feedback based on cognitive discrepancies between the FR (Facial Recognition) and FER (Facial Expression Recognition) datasets. This feedback is utilized to refine the parameters of the teacher network, with the feedback loss function $\mathcal{L}_f$ denoted by:

$$\mathcal{L}_f = f \cdot \mathrm{CE}(\hat{y}_{FR}, \mathcal{T}(x_{FR}; \theta_{\mathrm{t}})). \tag{5}$$

where the definition of the feedback coefficient f can be expressed as:

$$f = \eta_{\mathcal{S}} \cdot (\nabla_{\theta_{\mathcal{S}}^{(t+1)}} \mathrm{CE}(y_{FER}, \mathcal{S}(x_{FER}; \theta_{\mathcal{S}}^{(t+1)}))^{\top} \cdot \nabla_{\theta_{\mathcal{S}}} \mathrm{CE}(\hat{y}_{FR}, \mathcal{S}(x_{FR}; \theta_{\mathcal{S}}^{(t)}))). \quad (6)$$

The f is denoted as the dot product of two terms. First term: the gradient of the new student network on the debiased FER data. Second term: gradient of the old student network on biased FR data

3.3.3 Debiasing Mechanism

The performance of a novel student network on balanced Facial Expression Recognition (FER) data serves as the criterion for evaluation. More precisely, positive feedback coefficients are achieved (indicating a favourable $\mathcal{L}_f$ value) when both the student network operating on Face Recognition (FR) data and the novel student network focused on FER data share identical gradient orientations. This scenario promotes the modification of the teacher network through the employment of the gradient's current trajectory. Conversely, should the student network concerning FR data and the novel student network for FER data exhibit contrasting gradient orientations, the feedback coefficients assume a negative value. This acts as a deterrent to the teacher network's adjustment, advocating for the use of an opposing gradient direction. Thus, feedback operates as a crucial signal, retroactively influencing the teacher network by dictating the influence of its parameters on the student network's gradient for the extraction of unbiased features.

3.4. Temporal Encoder

The Transformer-based temporal encoder is a specialized neural network component that we have developed to specifically capture the temporal relationships between image features. Its purpose is to address the inherent feature bias in first-stage still images and enhance the robustness of dynamic expression recognition. This design leverages the self-attention mechanism of the Transformer architecture, which efficiently processes sequence data and learns temporal dependencies in image sequences. By applying self-attention in the temporal dimension, the temporal encoder allows the model to highlight keyframes and features within the image sequence, facilitating the extraction of spatio-temporal features. Moreover, the introduction of the temporal encoder enables the aggregation of information from neighbouring frames, resulting in a smoother overall feature representation for the video frame segments. This, in turn, leads to more robust dynamic expression analysis and prediction, enhancing the overall performance of the model.

3.5. Post-Process

Given that the Aff-Wild2 dataset is constructed from consecutive frames of videos, and considering that the formation of facial expressions unfolds over a period of time, it follows logically that significant alterations in facial expressions are unlikely to occur within a narrow sequence of adjacent frames. In light of this, we implemented a sliding window technique to refine the predictive outcomes, aiming to enhance the consistency of the expression labels. This process involves aggregating the frequencies of all predicted labels within each window. Subsequently, the label that predominates in frequency within a given window is designated as the representative expression for all frames encompassed by that window. By applying this sliding window approach across the entire dataset, we effectively facilitate the smoothing of predicted expression labels, thereby achieving a more coherent and accurate representation of facial expressions throughout the dataset.

4. Experiment and Results

In this section, we will provide a detailed description of the used datasets, the experiment setup, and the experimental results.

4.1. Datasets

FER Datasets. The 6th Workshop and Competition on Affective Behavior Analysis in-the-wild unveiled the Aff-wild2 database, a pivotal collection drawn from a range of studies[9, 10, 12–20, 38]. This resource was central to the EXPR Classification Challenge, featuring an audio-visual dataset comprised of 548 videos and approximately 2.7 million frames annotated for six fundamental facial expressions, as well as neutral and 'other' categories. To enrich the dataset's depth, we incorporated data from the AffectNet and ExpW databases; AffectNet contributed around 1 million facial images categorized into 11 expressions, while ExpW offered 91,793 images across seven expression categories. This integration aimed to enhance the comprehensiveness and utility of our analysis.

To further enhance the diversity of our dataset and bolster the generalizability of our model, we also randomly selected 8,000 images for each category from a merged dataset comprising AffecNet and ExpW. For images falling into the 'other' category, we relied exclusively on the Aff-Wild2 dataset, given the absence of such images in the alternate databases. The compilation is rounded off with the inclusion of the remaining images from these datasets as unlabeled samples, serving to augment our dataset's comprehensiveness.

FR Datasets. In our work on Face Recognition Datasets, we employ the MS1MV2 dataset[8] as the source of unlabeled data. This dataset, a semi-automatically refined it-

Table 1. Ablation study results on the validation set.

Method	Aff-Wild2	AffectNet and ExpW	MS1MV2	Temporal Encoder	Post-Process	F1 Score (%)
Baseline	✓					23.00
SSL	✓		✓			39.96
SSL	✓	✓	✓			40.57
SSL + Temporal	✓	✓	✓	✓		42.77
SSL + Temporal+ Post-process	✓	✓	✓	✓	✓	44.43

eration of the MS-Celeb-1M dataset[7], was developed by ArcFace[4] and contains approximately 85,000 identities and 5.8 million images. For the purposes of our experiments, we utilized a subset of the MS1MV2 dataset, curated from InsightFace[6], by uniformly selecting one-third of the images. This resulted in a comprehensive subset encompassing 1.94 million images.

4.2. Setup

All training facial images are identified and resized to 256×256 pixels, then further processed by random cropping to 224×224 pixels. The default architecture chosen for both the teacher and the student networks is ResNet50. Initially, learning rates are set at 1×10^{-2} for the teacher network and 1×10^{-3} for the student network. These rates are dynamically adjusted using a method known as cosine annealing. A batch size of 32 is utilized. The training process spans 100,000 steps and is conducted end-to-end on a single Nvidia V100 GPU. For temporal refinement, video inputs are segmented into 64-frame clips and undergo training for 20 epochs.

4.3. Metrics

We employ the average F1 Score as our evaluation metric, which is robust against class frequency variations and particularly suitable for imbalanced class distributions. The calculation of the average F1 Score is as follows:

$$F_1^c = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

$$F1 = \frac{1}{N}\sum_{c=1}^{N} F_1^c \tag{8}$$

where N represents the number of classes and c means c-th class.

4.4. Results

4.4.1 Test Set Performance

The performance of different teams in the Expression (Expr) Recognition Challenge on the official test set is shown in Table.2. Our method ranks third in the Aff-wild2 test set and achieves an F1 Score of 36.24%, which improves the F1 Score by 13.74% compared to the baseline model, proving the effectiveness of our method.

Table 2. The average F1 Score (in %) of different teams on the official Aff-wild2 test set.

Teams	F1 on Test Set
Netease Fuxi AI Lab [43]	50.05
CtyunAI [47]	36.36
USTC-IAT-United(Ours)	**36.24**
HSEmotion [29]	34.14
M2-Lab-Purdue [42]	32.28
KBS-DGU [24]	30.05
SUN CE [5]	28.77
AIOBT [27]	27.97
CAS-MAIS [35]	26.50
IMLAB [25]	22.96
baseline [20]	22.50

4.4.2 Ablation Study

To validate the efficacy of our approach, we executed ablation studies on each component and strategy within our method, with the outcomes presented in Table.1 It is evident that the application of the SSL (Semi-Supervised Learning) technique significantly enhances the recognition performance by 16.96%. Further augmentation of the facial expression dataset can lead to an increase in accuracy. Moreover, incorporating the temporal encoder results in an additional 2.2% improvement in accuracy, underscoring the importance of temporal learning. Ultimately, through post-processing, the model achieves an accuracy rate of 44.43%.

5. Conclusion

In this paper, we propose a two-phase approach to improve facial expression recognition. The first phase is called the spatial pre-training phase, in which, in order to address the problem of scarcity of facial expression data, we employ a semi-supervised learning technique to generate expression category pseudo labels for unlabeled facial data. At the same time, we uniformly sampled the labeled facial expression samples and implemented a debiased feedback learn-

ing strategy to address the problem of category imbalance in the dataset and possible data bias in semi-supervised learning. The second phase is the temporal refinement phase. In this phase, to compensate for the limitation and bias of obtaining features only from static images, we introduced a temporal encoder to learn and capture the temporal relationship between the features of neighbouring expression images to achieve more accurate dynamic facial expression recognition. Finally, our solution won the third place in Expression (Expr) Recognition Challenge, proving the effectiveness of our method.

Acknowledgement

This work was supported by the Natural Science Foundation of China (62276242), National Aviation Science Foundation (2022Z071078001), CAAI-Huawei MindSpore Open Fund (CAAIXSJLJJ-2021-016B, CAAIXSJLJJ-2022-001A), Anhui Province Key Research and Development Program (202104a05020007), USTC-IAT Application Sci. & Tech. Achievement Cultivation Program (JL06521001Y), Sci. & Tech. Innovation Special Zone (20-163-14-LZ-001-004-01).

References

[1] Jia-Ren Chang, Yong-Sheng Chen, and Wei-Chen Chiu. Learning facial representations from the cycle-consistency of face. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9680–9689, 2021. 1

[2] Yunliang Chen and Jungseock Joo. Understanding and mitigating annotation bias in facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14980–14991, 2021. 1, 2

[3] 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. 4

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 6

[5] Denis Dresvyanskiy, Maxim Markitantov, Jiawei Yu, Peitong Li, Heysem Kaya, and Alexey Karpov. Sun team's contribution to abaw 2024 competition: Audio-visual valence-arousal estimation and expression recognition, 2024. 6

[6] Jia Guo. Insightface: 2d and 3d face analysis project. https://github.com/deepinsight/insightface, Accessed on Month Day, Year. 6

[7] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016. 6

[8] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 5

[9] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2328–2336, 2022. 1, 5

[10] Dimitrios Kollias. Abaw: Learning from synthetic data & multi-task learning challenges. In *European Conference on Computer Vision*, pages 157–172. Springer, 2023. 5

[11] Dimitrios Kollias. Multi-label compound expression recognition: C-expr database & network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5598, 2023.

[12] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 5

[13] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021.

[14] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021.

[15] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019.

[16] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019.

[17] Dimitrios Kollias, Attila Schulc, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 637–643. IEEE, 2020.

[18] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021.

[19] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. *arXiv preprint arXiv:2303.01498*, 2023.

[20] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Stefanos Zafeiriou, Irene Kotsia, Alice Baird, Chris Gagne, Chunchang Shao, and Guanyu Hu. The 6th affective behavior analysis in-the-wild (abaw) competition, 2024. 1, 5, 6

[21] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 2

[22] Hangyu Li, Nannan Wang, Xi Yang, Xiaoyu Wang, and Xinbo Gao. Towards semi-supervised deep facial expression recognition with an adaptive confidence margin. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4166–4175, 2022. 2

[23] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017. 1, 2

[24] Li Lin, Sarah Papabathini, Xin Wang, and Shu Hu. Robust light-weight facial affective behavior recognition with clip, 2024. 6

[25] Seongjae Min, Junseok Yang, Sangjun Lim, Junyong Lee, Sangwon Lee, and Sejoon Lim. Emotion recognition using transformers with masked learning, 2024. 6

[26] 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

[27] Bach Nguyen-Xuan, Thien Nguyen-Hoang, and Nhu Tai-Do. Emotic masked autoencoder with attention fusion for facial expression recognition, 2024. 6

[28] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016. 2

[29] Andrey V. Savchenko. Hsemotion team at the 6th abaw competition: Facial expressions, valence-arousal and emotion intensity prediction, 2024. 6

[30] 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*, pages 6248–6257, 2021. 2

[31] 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. *Advances in neural information processing systems*, 33:596–608, 2020. 2

[32] Yingli Tian, Takeo Kanade, and Jeffrey F Cohn. Facial expression recognition. *Handbook of face recognition*, pages 487–519, 2011. 1

[33] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, et al. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022. 3

[34] Zhengyao Wen, Wenzhong Lin, Tao Wang, and Ge Xu. Distract your attention: Multi-head cross attention network for facial expression recognition. *arXiv preprint arXiv:2109.07270*, 2021. 2

[35] Zhuofan Wen, Fengyu Zhang, Siyuan Zhang, Haiyang Sun, Mingyu Xu, Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Multimodal fusion with pre-trained model features in affective behaviour analysis in-the-wild, 2024. 6

[36] Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. Rethinking infonce: How many negative samples do you need? *arXiv preprint arXiv:2105.13003*, 2021. 2

[37] Zhiding Yu and Cha Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 435–442, 2015. 1

[38] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017. 1, 5

[39] Dan Zeng, Zhiyuan Lin, Xiao Yan, Yuting Liu, Fei Wang, and Bo Tang. Face2exp: Combating data biases for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20291–20300, 2022. 2

[40] 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

[41] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021. 3

[42] Wei Zhang, Bowen Ma, Feng Qiu, and Yu Ding. Facial affective analysis based on mae and multi-modal information for 5th abaw competition. *arXiv preprint arXiv:2303.10849*, 2023. 6

[43] Wei Zhang, Feng Qiu, Chen Liu, Lincheng Li, Heming Du, Tiancheng Guo, and Xin Yu. Affective behaviour analysis via integrating multi-modal knowledge, 2024. 6

[44] Xiaoqin Zhang, Min Li, Sheng Lin, Hang Xu, and Guobao Xiao. Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2

[45] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 418–434. Springer, 2022. 2

[46] Yuhang Zhang, Yaqi Li, Xuannan Liu, Weihong Deng, et al. Leave no stone unturned: Mine extra knowledge for imbalanced facial expression recognition. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[47] Weiwei Zhou, Jiada Lu, Chenkun Ling, Weifeng Wang, and Shaowei Liu. Boosting continuous emotion recognition with self-pretraining using masked autoencoders, temporal convolutional networks, and transformers, 2024. 6