Facial Expression Recognition Method Based on Dynamic Weighted Attention and Noise Label Optimization

YiHeng Sun

Abstract— **Facial expression, as the core carrier of human emotion transmission, has important value in the fields of interpersonal interaction, mental health assessment and intelligent human-machine system. However, the existing facial expression recognition methods still face challenges such as high inter-class similarity, significant intra-class differences, prominent scale sensitivity, and serious noise interference. Traditional methods rely on artificial feature extraction, which has the defect of insufficient generalization ability. Although the method based on deep learning has a breakthrough in performance, the problems of insufficient global information capture and weak ability to suppress uncertain labels limit its application in complex scenarios. Therefore, this paper proposes a facial expression recognition model based on dynamic weighted attention and noise label optimization, aiming to construct an efficient and robust recognition framework through multi-modal feature fusion and adaptive optimization strategy. In this paper, extensive experiments are carried out on the RAF-DB dataset. The experimental results show that the accuracy of the model on the dataset with test noise ratios of 10 %, 20 % and 30 % reaches 89.97 %, 89.13 % and 87.22 %, respectively. The accuracy is better than the current state-of-the-art expression recognition model.**

Index Terms— **Facial expression recognition, deep learning, attention mechanism, noisy label learning, distribution learning.**

I. INTRODUCTION

As an important non-verbal communication method, facial expression carries rich emotional information and is an important manifestation of human emotions and psychological states. Facial expression recognition is becoming more and more widely used in various industries. However, the existing facial expression recognition methods still face challenges such as high inter-class similarity, significant intra-class differences, and prominent scale sensitivity. In addition, due to the subjectivity of data set labeling, the labels in the data set may have noise, which further aggravates the above problems.

In view of these problems, an enhanced deep 3D convolutional neural network was proposed for facial expression recognition in the early stage. The deep learning method of adding facial landmarks with image features was explored to improve the attention to details and help to promote the invariance of intra-class differences, that is, a set of key points of face images were extracted as facial landmarks to provide sparse representation of key facial regions and supplement direct image features. Although the combination of facial feature points and image features reduces intra-class differences and inter-class similarities to a certain extent, only connection is not enough to explore the correlation between feature points. In order to deal with the problem of inconsistent data quality caused by subjectivity in the labeling process, label distribution learning is proposed. As a more expressive method than traditional discrete labels, it can effectively describe the ambiguity in expressions. LDL solves the problem of inter-class similarity by learning the probability distribution of different expressions, and the intra-class difference can be effectively suppressed by the average expression anchoring module. Based on this, Wang et al. [1] proposed SCN to suppress the uncertainty of the real scene FER, and used the attention mechanism to weight each training face sample. This method estimates the importance weight of each image to determine whether its label is affected by noise, and gives the score. However, these methods have obvious limitations. These methods usually only consider the information of a single sample or a small batch of samples, ignoring the use of global information. This limitation causes the model to make unreliable decisions when dealing with noise labels, and some clean data is not fully utilized, while some noise data interferes with the learning process of the model.

In order to better deal with these problems, this study uses the dynamic weighted multi-head cross-attention mechanism proposed (DWMCA) in this paper. Through the channel weight adaptive and multi-head cross-attention mechanism, the cross-modal feature expression ability is enhanced and the cross-modal feature interaction is realized. For the problem of uncertainty in annotation data, an adaptive label cleaning model (APCL) is designed. The model introduces the method of reliability balance, and sets learning anchors in the embedding space with label distribution and multi-head self-attention mechanism to optimize the performance of correct label prediction through reliability balance. This method can not only better deal with the expression recognition task in complex scenes, but also effectively solve the problem of noise labeling, so as to ensure that the model can maintain good recognition ability in high noise environment.

The main contributions of this study can be summarized as follows :

1.The dual-stream architecture is introduced, and the

Manuscript received April 18, 2025
YiHeng Sun, School of software, Tiangong University, Tianjin, China

DWMCA method is proposed by combining image stream and surface stream. The dynamic channel weighting and cross-attention mechanism strategy are adopted to adaptively enhance the response of key regions, and the gating mechanism is introduced to control the fusion intensity of cross-modal information.

2.The adaptive label purification model (APCL) is designed. The method is mainly composed of two modules : basic label prediction and label balance purification. The main label distribution and confidence value of the two modules are calculated, and finally the corrected label distribution is obtained by weighted fusion.

3.The method of this paper has been widely evaluated on the RAF-DB dataset. The experimental results show that the method shows superior performance on the RAF-DB dataset with 10 %, 20 % and 30 % noise ratio.

II. Procedure for Paper Submission

A. Facial Expression Recognition

With the rapid development of artificial intelligence and computer vision technology, facial expression recognition technology has gradually become a research hotspot in the fields of human-computer interaction, emotional computing and intelligent medical treatment. As an important non-verbal communication method, facial expression carries rich emotional information and is an important manifestation of human emotions and psychological states[20]. Therefore, how to automatically recognize and understand facial expressions through computers has become an important topic in the field of artificial intelligence.

With the rapid development of deep learning technology in computer vision tasks, deep learning solutions are increasingly used to deal with challenging FER tasks and achieve promising performance. Wang et al. [2] proposed a regional attention network (RAN) to capture facial regions for occlusion and posture change FER. Farzaneh and Qi [3] introduced a deep attention center loss (DACL) method to estimate the attention weight of features to enhance discrimination. The sparse center loss is designed to achieve intra-class compactness and inter-class separation. Wang et al. [4] proposed a self-correcting network (SCN) to suppress uncertainty and prevent the network from overfitting mislabeled samples. Shi et al. [5] designed a modified representation module (ARM), which can reduce the weight of eroded features and decompose face features to simplify representation learning.

For the convolutional neural network model, the loss function plays a crucial role, which can measure the gap between the prediction and the actual data[21]. In recent years, many studies on loss functions have further improved the recognition ability of FER. Hadsell et al.proposed Contrastive Loss [6] for extracting deep features to bring together features with similar labels and separate features with different labels. Inspired by Center Loss [7], Li et al.embedded Locality Preserving Loss (LP-Loss) into DLP-CNN, and performed local clustering on deep features through k-nearest neighbor algorithm to enhance intra-class compactness. Cai et al. [8] improved the central loss by adding an additional objective function called Island Loss to achieve both intra-class compactness and inter-class separation. The island loss maximizes the cosine distance between the class centers in the embedding space.Li et al. [9] proposed the cosine version of the center loss and the island loss, which can maximize the cosine similarity between the features belonging to one class in the intra-frame loss and the inter-frame loss, and minimize the cosine similarity between the class centers in the embedded space. Farzaneh et al. [10] proposed a distribution-independent loss (DDA Loss), which allows the model to implicitly perform inter-category separation between majority and minority classes in the case of extreme class imbalance. DDA Loss adjusts the Euclidean distance of samples between all classes in the embedding space during forward propagation[22].

B. Noisy Label Learning

Although the traditional facial expression recognition method performs well on some small-scale datasets, the effect in the real scene dataset is not satisfactory. In the process of collecting and sorting out the real scene facial expression recognition data set, due to the influence of various interference factors, such as the interference of different races, ages, genders and other semantic levels, these data sets will contain noise labels due to these interference in the expression image and the different cognition of the annotators when using crowdsourcing annotation[23]. Image noise labels will seriously affect the learning of deep learning models and reduce their robustness. Therefore, solving these problems is crucial for facial expression recognition algorithms.

Wang et al. proposed SCN to suppress the uncertainty of FER in real scenes, and made full use of the attention mechanism to weight each training face sample. The SNEFER model proposed by Gao et al. [11] effectively solves the negative impact of noise labels on the FER model through its unique contrast regularization branch and BCR loss. It does not need to manually select noise samples, and only returns positive gradients from clean samples, so that the FER model remains robust at different noise levels. Zhang et al. [12] proposed an innovative uncertainty learning method, called relative uncertainty learning (RUL). RUL does not assume the Gaussian uncertainty distribution of all data sets, but constructs an additional branch to learn uncertainty from the relative difficulty of the sample through feature mixing. Specifically, uncertainty is used as a weight to mix facial features, and an additive loss is designed to encourage uncertainty learning. She et al. [13] proposed a model DMUE to solve the ambiguity of annotations from two perspectives : potential distribution mining and pairwise uncertainty estimation.For the former, an auxiliary multi-branch learning framework is introduced to better mine and describe the potential distribution in the label space. For the latter, we make full use of the pairwise relationship of semantic features between instances to estimate the degree of ambiguity in the instance space. Zhang et al. [14] proposed a new method called ' Erasing Attention Consistency (EAC) ' to deal with the problem of noise labels in facial expression recognition (FER). A new EAC method is proposed to deal with the facial expression recognition (FER) problem with noise labels by automatically suppressing noise samples during training. This method uses the flipping semantic consistency of facial images to design an unbalanced framework[24, 25, 26], and prevents the model from paying attention to only part of the features by randomly erasing the input image, effectively preventing the model from remembering noise samples.

III. MATH

In this paper, a dynamic weighted cross-attention mechanism with two-stream structure is proposed to solve the intra-class difference inter-class similarity and multi-scale problems in FER, and an adaptive label purification module is designed to solve the problems caused by noise labels, as shown in Figure 1. In this section, we will first introduce the overall structure of the network. Then, the dynamic weighted cross-attention module and the noise label processing module APCL are explained in detail. Finally, the multi-loss joint training optimizes the entire framework.

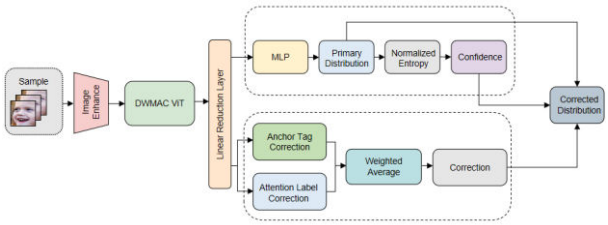

Figure 1 Network architecture

A. *Dynamic weighted cross attention module*

In order to overcome the problem of insufficient feature extraction and irrelevant context semantic information in facial expression recognition tasks, the cross-attention mechanism module will be improved. The extraction module generates dynamic weight vectors for image flow and key point flow respectively before cross-attention calculation by designing a dynamic channel weighting mechanism. The weight is adaptively adjusted according to input features by lightweight MLP, and multi-head cross-attention enhancement is adopted. The single cross-attention is extended to a multi-head mechanism, each head independently learns the feature interaction of different subspaces, and the gating mechanism is introduced to control the fusion intensity of cross-modal information to avoid noise interference, so as to ensure the reasonable and effective use of the correlation between modes and the feature information between different modes to achieve higher accuracy.

B. *Dynamic channel weighting*

Through a lightweight MLP module, a dynamic weight is calculated for each feature channel, and the weight generation is based on the quality and importance of the current input feature. According to the different input features, the information quality of facial landmarks and image regions is adjusted. Through the automatic learning of the training process, the role of important channels is enhanced during feature fusion, and the influence of noise and irrelevant information is suppressed.

Firstly, each channel of the input feature is globally averaged pooled, and the spatial information is compressed to obtain the channel-level statistics.

$$g = GAP(X_{img}) \in R^D$$

This step reduces the feature dimension to capture the global importance of each channel. Then it will be mapped to the low latitude space H to reduce the number of parameters.

$$h = ReLU(W_1 \cdot g + b_1), W_1 \in R^{D\times H}, b_1 \in R^H$$

The hidden layer feature h is mapped back to the original channel dimension, and the weight is generated by Sigmoid. The formula is as follows:

$$\alpha_{img} = Sigmoid(W_2 \cdot h + b_2), W_2 \in R^{H\times D}, b_2 \in R^D$$

The output represents the importance weight of each channel, and the final weighted feature is as flow:

$$\tilde{X}_{img} = X_{img} \odot \alpha_{img}, \tilde{X}_{lm} = X_{lm} \odot \alpha_{lm}$$

C. *Multihead cross-attention enhancement*

The single cross-attention mechanism is extended to a multi-head mechanism, each head independently learns the feature interaction of different subspaces. For the MSA in the image stream, the input is mapped to three image matrices, query matrices, key matrices and value matrices through three linear transformations. For the MSA in the surface stream, the input is mapped to three image matrices, query matrices, key matrices and value matrices through three linear transformations.

After obtaining the weight processed by the dynamic channel weighting, three linear transformations are performed, followed by the calculation of single-head attention, followed by multi-head splicing.

$$DW - MCA = Concat(Head_1, \ldots, Head_h)W^0$$

$W_i^Q, W_i^K, W_i^V \in R^{D\times d}$ represents the matrix projection of each projection, $W^0 \in R^{h.d\times D}$ represents the output projection matrix.

D. *Adaptive label purification model*

The designed APCL model is mainly composed of two modules : fine-tuning BLP module and label balance purification LBP module.

Basic Tag Prediction BLP module: In this module, the comprehensive feature vector extracted from the DWMCA weighted cross-attention module is used as input, which provides correctness and reliability in the final output vector with a length of 768. An additional linear reduction layer is used to reduce the feature vector size to 128. The logits generated by the multi-layer perceptron are used to generate the main label distribution. The multi-layer perceptron MLP includes various hidden layers, enabling them to process information with extremely high accuracy and accuracy. Finally, the normalized entropy is used to calculate the confidence value based on the main label distribution to evaluate the reliability of these models.

Label balance purification LBP module: This module mainly designs two methods of anchor label correction and attention correction, and makes the final prediction label accurate and stable by using geometric similarity and advanced multi-head attention mechanism, which ensures that the model is resilient even in the case of noise or insufficient data, and can make credible predictions even in the case of ambiguity. The possible bias and overfitting are minimized.

Anchor label correction : Anchor points $a^{i,j}$ ($i \in \{1,2 \ldots, N\}$, $j \in \{1,2 \ldots K\}$) are defined as points in the embedded space. Let $\mathcal{A}$ be the set of all anchor points. During training, using K trainable anchors for each label, where k is a hyperparameter. Another label distribution $m^{i,j} \in \mathcal{P}^N$ is assigned to the anchor $a^{i,j}$, $m^{i,j}$ is defined as follows :

$$m_k^{i,j} = \begin{cases} 1, \text{if } k = i \\ 0, \text{otherwise} \end{cases}$$

Calculate geometric distance and similarity : In order to correct the final label and stabilize the distribution, the geometric information of the similarity between the embedding vectors and the fixed number of learnable points in the embedding space are used, which are called anchor points. Similarity score $s^{ij}(e)$ is the normalized measure of

similarity between embedding x and anchor point a^{ij}. The distance between embedding e and anchor a for each batch and category is defined as:

$$d(e,a) = \sqrt{\sum_{dim_e} |a-e|^2}$$

Here, δ is a hyperparameter for Softmax calculations to control the steepness of the function. The default value of δ is 1.0. According to the similarity score, the anchor label correction term is obtained.

$$t_g(e) = \sum_{i}^{N} \sum_{j}^{K} s^{ij}(e) m^{ij}$$

Multi-head self-attention mechanism assigns label confidence scores : for multi-head attention, Let the query embedding be $q \in \mathcal{R}^{d_Q}$, the key embedding be $k \in \mathcal{R}^{d_K}$, and the value embedding $v \in \mathcal{R}^{d_V}$, with the help of independent learning projection, they can be modified with h, which is the center of attention. These parameters are then provided to the attention pooling layer. Finally, these outputs are changed and another linear projection integral is used.

In order to further correct and stabilize the label distribution, an attention-based similarity function is used. Embed x through the multi-headed self-attention layer to obtain the attention correction term t_a:

$$t_a = \text{softmax}(W_{out})$$

Finally, in order to fuse the overall label correction term, a weighted sum is used, where the weighting is controlled by the confidence of the label correction, which is expressed as :

$$t = \frac{c_g}{c_g + c_a} t_g + \frac{c_a}{c_g + c_a} t_a$$

The label with the maximum value in the final correction label distribution L_{final} is used as the correction label or the final prediction label.

E. Joint loss training

The loss function used to train the model consists of three terms, negative log-likelihood loss, anchor loss and center loss.

In order to ensure that each batch label is correctly classified, we use the negative log-likelihood loss between the corrected label distribution Li and the label yi as the class distribution loss $\mathcal{L}_{cls}$.

$$\mathcal{L}_{cls} = -\sum_{i}^{m} \sum_{j}^{N} y_j^i \log L_j^i$$

In order to maximize the Euclidean distance between anchor points of different categories, the discriminant of embedding space is enhanced. By optimizing the distribution of anchor points, the feature representations of different categories are separated as much as possible in the embedding space.

$$\mathcal{L}_a = -\sum_{i} \sum_{j} \sum_{k} \sum_{l} \left| a^{ij} - a^{kl} \right|_2^2$$

In this paper, the center loss is designed to enhance the intra-class compactness of the feature space by optimizing the distance between the sample embedding and the anchors, and to achieve the goal of ' intra-class aggregation and inter-class separation ' in collaboration with the anchor loss.

$$\mathcal{L}_c = \min_k \left\| x^i - a^{y^i,k} \right\|_2^2$$

The final loss function can be defined as the following formula :

$$\mathcal{L}_{total} = \lambda_{cls} \mathcal{L}_{cls} + \lambda_a \mathcal{L}_a + \lambda_c \mathcal{L}_c$$

λ_{cls} , λ_a, λ_c are hyperparameter that preserves the loss function on the same scale.

IV. Experiments

A. Datasets

RAF-DB [15] (Real-World Affective Face Database) is a real-world dataset widely used in facial expression recognition research. The dataset is annotated by 40 trained manual annotators and contains 15,339 facial images, covering six basic expressions (happiness, surprise, sadness, anger, disgust, fear) and neutral expressions. These images are derived from real scenes in daily life, aiming to provide researchers with a challenging environment to improve the robustness of facial expression recognition algorithms.

In the experiment of this paper, 12,271 images were selected for training and 3,068 images were used for testing. Through this design, the performance of the model in the real world scene can be better evaluated. In addition, the overall sample accuracy of the test set is used as an important indicator to measure the performance of the model.

RAF-DB not only provides a wealth of facial expression samples, but also reflects the diversity of human emotions in daily life, which provides valuable data support for the development and verification of facial expression recognition technology.

B. Implementation Details

For each data set, get cropped and aligned images. Their sizes were adjusted to 256 × 256, and 224 × 224 were randomly intercepted. In order to deal with overfitting and data imbalance in specific expression categories, image-assisted enhancement method is used to preprocess the data. 512 images are evenly sampled from each face data, covering multiple angles, ensuring the diversity of captured expressions, merging all face images into a unified data set, and disrupting the order to eliminate source correlation. During training, 500 images are obtained for each class category from the collection. The batch size of the dataset is set to 32. Using the Adam optimizer learning rate scheduler, the learning rate is initialized to 0.001. Set gamma to 0.9 to attenuate the learning rate after each epoch. The training process ends at the 200th epoch. In this paper, NVIDIA RTX 3090 is used to train the model in the environment of CUDA 11.3, PyTorch 1.12.0, torchvision 0.13 and Python 3.9.

C. Ablation study

In order to evaluate the effect of different modules respectively, this paper conducts ablation research on RAF-DB at 30 % noise level based on IR50 and MobileFaceNet benchmark models. In order to obtain more realistic performance, this paper uses the average of the last five iteration results, as shown in Table 1. For this reason, some conclusions can be drawn. The accuracy of using the DWMCA module on the benchmark model is increased by 7.14 %, which proves the effectiveness of the DWMCA module. Similarly, on this basis, the APCL module is added, and the accuracy rate is increased by 8.65 %, which proves that the APCL module can reduce the impact of noise on the model and avoid overfitting noise data. In summary, the

experimental results can well prove the effectiveness of the model.

Table I Ablation experiment

BaseLine	DWMCA	APCL	30% RAF-DB
√			71.43
√	√		78.57
√	√	√	87.22

D. Performance Comparison

This paper quantitatively evaluates the improvement effect of the proposed model compared with other state-of-the-art noise label facial expression recognition (FER) methods. The robustness of the model is explored by adding three noise ratios of 10 %, 20 % and 30 % to the RAF-DB dataset for a fair comparison. As shown in Table 2, the proposed method is much better than other state-of-the-art FER noise label learning methods. For example, the performance of the proposed method on RAF-DB is 7.79 %, 9.03 % and 9.76 % higher than that of the SCN method, respectively.

Table II Contrast experiment on 10% Noise RAF-DB

Methods	**Noise**	**RAF-DB**
SCN [12]	10	82.18
DMUE [16]	10	83.19
RUL [17]	10	86.22
EAC [18]	10	88.02
Ada-DF[19]	10	87.81
Our	10	89.97

Table 3 : Contrast experiment on 20% Noise RAF-DB

Methods	**Noise**	**RAF-DB**
SCN [12]	20	80.10
DMUE [16]	20	81.02
RUL [17]	20	84.34
EAC [18]	20	86.05
Ada-DF[19]	20	86.67
Our	**20**	**89.13**

Table 4 : Contrast experiment on 30% Noise RAF-DB

Methods	**Noise**	**RAF-DB**
SCN [12]	30	77.46
DMUE [16]	30	79.41
RUL [17]	30	82.06
EAC [18]	30	84.42
Ada-DF[19]	30	84.38
Our	**30**	**87.22**

V. CONCLUSION

The facial expression recognition method proposed in this paper combines the dynamic weighted multi-head cross-attention mechanism (DWMCA). Through the combination of dynamic channel weight adjustment and multi-head attention mechanism, the cross-modal interaction ability of image stream and landmark stream features is enhanced. With the adaptive label purification model (APCL), the label distribution is optimized by anchor correction and multi-head attention mechanism, which solves the problems of intra-class difference, inter-class similarity and label noise in facial expression recognition, so as to improve the performance of facial expression recognition model. Extensive experiments on RAF-DB datasets. It shows the effectiveness and robustness of the proposed method to enhance the robustness of the model to noise labels. In the future work, we intend to integrate more FER-related tasks, such as combining with semi-supervised learning, using unlabeled data to enhance the reliability of noise label correction, reducing the dependence on labeled data, using dynamic noise estimation, designing online noise detection mechanism, evaluating the label quality of training samples in real time and dynamically adjusting the learning weight. These efforts will help promote the development of the FER field and expand the capabilities of our proposed methods.

REFERENCES

[1] Wang K, Peng X, Yang J, et al. Region attention networks for pose and occlusion robust facial expression recognition[J]. IEEE Transactions on Image Processing,2020,29:4057-4069.

[2] Meng, D., Qiao, Y., Wang, K., Peng, X., Yang, J.: Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing 29, 4057–4069 (2020)

[3] Meng, D., Qiao, Y., Wang, K., Peng, X., Yang, J.: Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing 29, 4057–4069 (2020)

[4] Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large scale facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6897–6906 (2020)

[5] Shi, J., Zhu, S., Liang, Z.: Learning to amend facial expression representation via de-albino and affinity. arXiv preprint arXiv:2103.10189 (2021)

[6] Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping[C]. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition . IEEE, 2006, 2: 1735-1742.

[7] Wen Y, Zhang K, Li Z, et al. A discriminative feature learning approach for deep face recognition[C]. Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14. Springer International Publishing, 2016: 499-515.

[8] Cai J, Meng Z, Khan A S, et al. Island loss for learning discriminative features in facial expression recognition[C]. 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition. IEEE, 2018: 302-309.

[9] Sohn K. Improved deep metric learning with multi-class n-pair loss objective[J]. Advances in Neural Information Processing Systems, 2016, 29.

[10] Farzaneh A H, Qi X. Discriminant distribution-agnostic loss for facial expression recognition in the wild[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020: 406-407.

[11] Gao Y,Ren W,Wang Q,et al.SNEFER: Stopping the Negative Effect of Noisy Labels Adaptively in Facial Expression Recognition[J].IEEE Sensors Journal, 2024,24(11):8622-18632.

[12] Zhang Y, Wang C, Deng W. Relative uncertainty learning for facial expression recognition[J]. Advances in Neural Information Processing Systems, 2021, 34: 17616-17627.

[13] Wang K, Peng X, Yang J,et al. Suppressing Uncertainties for Large-Scale Facial Expression Recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2020:6897-6906.

[14] She J, Hu Y,Shi H ,et al. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.2021:6248-6257.

[15] Li, S., Deng, W., Du, J., & Zhang, Z. (2017). Reliable crowd-sourcing and deep locality preserving learning for expression recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 285-2861).

[16] She, J., Hu, Y., Shi, H., Wang, J., Shen, Q., & Mei, T. (2021). 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* (pp. 6248-6257).

[17] Zhang, Y., Wang, C., & Deng, W. (2021). Relative uncertainty learning for facial expression recognition. *Advances in Neural Information Processing Systems*, 34, 17616-17627.

[18] Y. Zhang, C. Wang, X. Ling, W. Deng, Learn from all: Erasing attention consistency for noisy label facial expression recognition, in: European Conference on Computer Vision, Springer, 2022, pp. 418–434.
[19] Liu, S., Xu, Y., Wan, T., & Kui, X. (2023). A dual-branch adaptive distribution fusion framework for real-world facial expression recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.
[20] Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A.,Bengio, Y., et al.: A closer look at memorization in deep networks. In: ICML (2017)
[21] Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2011). Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *Proceedings of the International Conference on Computer Vision* (pp. 2106-2112).
[22] Han, B., Yao, J., Niu, G., Zhou, M., Tsang, I., Zhang, Y., Sugiyama, M.: Masking: A new perspective of noisy supervision. In: NIPS (2018)
[23] Li, S., Deng, W., Du, J., & Zhang, Z. (2017). Reliable crowd-sourcing and deep locality preserving learning for expression recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 285-2861).
[24] Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L.: Mentornet: Learning data driven curriculum for very deep neural networks on corrupted labels. In: ICML (2018)
[25] Mollahosseini, A., Hasani, B., & Mahoor, M.H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1), 18-31.
[26] Ruan, D., Mo, R., Yan, Y., Chen, S., Xue, J., & Wang, H. (2022). Adaptive deep disturbance disentangled learning for facial expression recognition. *International Journal of Computer Vision*, 1-23.