A Dual-Branch Facial Expression Recognition Network with Noisy Label Learning and Identity Invariance

HuaYu Liu

***Abstract*—Facial Expression Recognition (FER) plays a significant role in daily life. However, prevalent challenges in expression datasets - including label noise, substantial intra-class variations, and small inter-class differences - severely impact model performance. To address these issues, this paper proposes a dual-branch facial expression recognition network framework. The framework models the similarity distribution of expression data through data modeling and introduces label distribution correction to handle label noise. To mitigate facial identity interference in expression recognition that causes large intra-class variations and small inter-class differences, an average expression anchoring module is employed. Extensive experiments conducted on the RAF-DB dataset demonstrate the model's superior performance, achieving accuracies of 89.09%, 88.62%, and 86.27% on datasets with test noise ratios of 10%, 20%, and 30% respectively, outperforming current state-of-the-art facial expression recognition models in accuracy.**

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

I. Introduction

Facial Expression Recognition (FER) serves as a vital tool for identifying and classifying human emotional states. Its core task involves extracting feature representations from facial images and categorizing them into specific expression classes. With extensive applications in psychology, medical research, user experience studies, and human-computer interaction, the accuracy and robustness of FER technology are critically important. However, existing FER methods face significant challenges in real-world complex scenarios that often compromise model performance. Specifically, the critical obstacles hindering FER model improvement are the high inter-class similarity (subtle distinctions between different expressions) and substantial intra-class variation (diverse manifestations within the same expression category) prevalent in expression datasets. Furthermore, potential label noise caused by annotator subjectivity may further exacerbate these issues.

Currently, many FER datasets such as RAF-DB [8], SFEW/AFEW [6], FERPlus [1], and AffectNet [10] use discrete classification models for annotation, but this approach struggles to accurately capture the complexity and ambiguity of facial expressions, especially in "wild" environments.the boundaries between different expression categories are often unclear, leading to high inter-class similarity. Meanwhile, the same expression can vary significantly among different individuals, resulting in high intra-class variability. This ambiguity and variability make the expression recognition task more challenging and increase the difficulty for models to learn. Additionally, the subjectivity in the annotation process leads to inconsistent data quality, further complicating the stable performance of expression recognition systems. To address this issue, Label Distribution Learning (LDL) has been proposed as a more expressive method than traditional discrete labels, capable of effectively describing the ambiguity in expressions. LDL resolves the problem of inter-class similarity by learning the probability distribution of different expressions, while the average expression anchoring module effectively suppresses intra-class variability.

Meanwhile, existing FER methods addressing label noise typically rely on an importance learning branch that estimates a weight for each image to assess potential label corruption [19,13,23]. However, these approaches exhibit notable limitations: 1) The strong learning capacity of Deep Neural Networks (DNNs) [23] makes the importance learning branch prone to overfitting noisy patterns; 2) They predominantly leverage information from individual samples or mini-batches while neglecting global data characteristics. Such limitations may lead to unreliable noise judgment underutilization of clean data and insufficient suppression of noisy samples during training.

To better resolve these issues, this study proposes a similarity distribution-based modeling approach that holistically integrates noise-corrupted and clean label distributions. By explicitly modeling discrepancies between predicted values and target labels, the method achieves more reliable noise detection. Building upon this foundation, Label Distribution Learning (LDL) is employed to characterize expression ambiguity, while the Average Expression Anchoring Module (terminology consistency ensured) suppresses high inter-class similarity and intra-class variation through feature space regularization. This framework not only enhances performance in complex real-world scenarios but also systematically addresses label noise, ensuring robust recognition capabilities even under high-noise conditions.

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

1. the average expression anchoring module is used to effectively solve the problems of high inter-class similarity and intra-class difference in FER.
2. A noise-handling module (comprising noise modeling and label distribution correction) is employed to address dataset labeling errors.

Manuscript received April 15, 2025
HuaYu Liu, School of computer science and technology, Tiangong University, Tianjin, China.

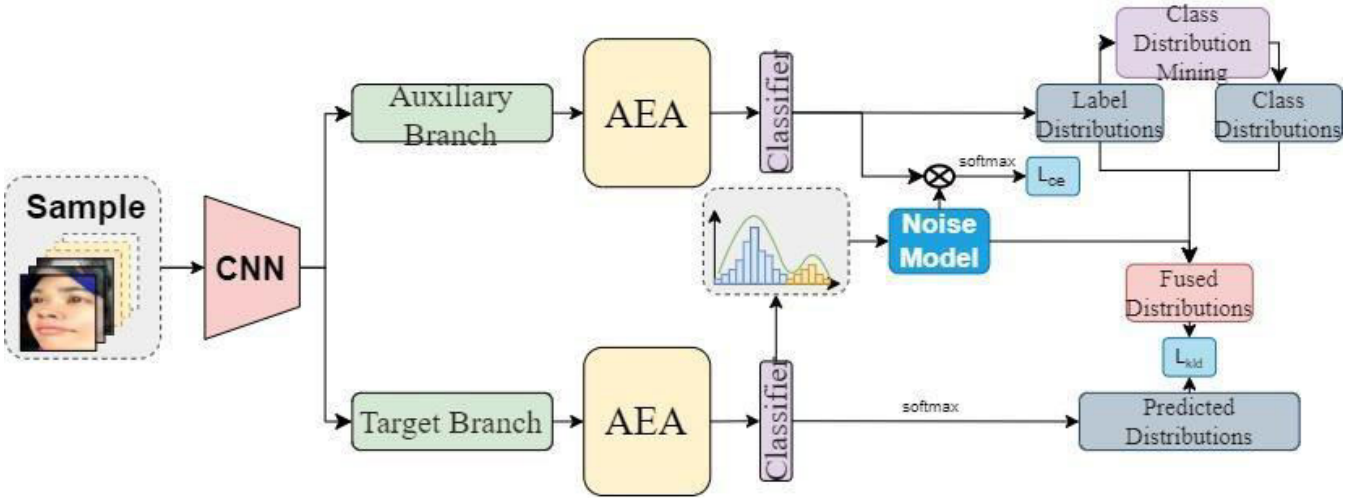

Fig 1 Network architecture

3. Extensive experiments conducted on the RAF-DB dataset demonstrate the method's superior performance under 10%, 20%, and 30% noise ratios.

II. METHOD

This paper proposes a dual-branch network structure to address the label noise problem in FER and the impact of facial features on FER, as shown in Fig 1. In this section, the overall structure of the network will be introduced first. Then, the average expression description module and the noise label processing module will be explained in detail. Finally, the entire framework is optimized through multi-loss joint training.

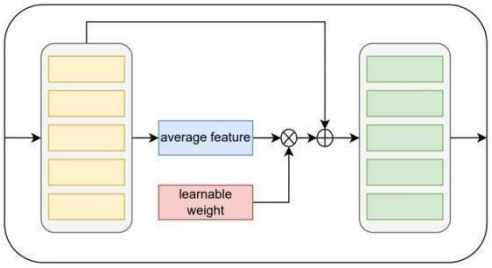

Fig 2 Average Expression Anchoring

A. Average Expression Anchoring

To enhance the performance of facial expression recognition and alleviate the impact of high inter-class similarity and intra-class diversity, an average expression description module is introduced. This module aims to reduce the influence of individual facial features on expression recognition, thereby improving the model's ability to focus on expression-related information.

Generally, the feature vectors obtained through the branch structure are denoted as f_i. The average expression anchoring module is composed of learnable attention weights α and the average face features obtained from a batch of expression features, and the element-wise multiplication and addition are performed, as shown in the following formula:

$$\bar{} = {}^{1}\mathbf{I}_{=1} \quad (1)$$

$$= * \bar{} + \quad (2)$$

Wherein, b is the size of a batch.

B. Beta Noise Modeling Module

To obtain the probability that each sample is noise, this paper employs a beta noise modeling module. By modeling the entire dataset and then using cosine similarity through the model, the probability that each data sample has a clean label is obtained.

Suppose the distribution obtained through the classifier is $x_i \in R^C$, and the cosine similarity S_i of x_i is calculated as follows:

$$= \frac{\mathbf{I}_{=1} \cdot}{\sqrt{\mathbf{I}_{=1}\mathbf{I}_{=1}\,0^{\,2}}} = \cos(\cdot) \quad (3)$$

C is the number of categories.

Mixture model is a widely used unsupervised modeling technique defined as:

$$p(S) = \mathbf{I}^{K}_{m=1}\, \sigma_m\, p(S|m) \quad (4)$$

Model the similarity distribution using a mixed beta model, and then use the generated noise model to provide the probability that each sample belongs to a clean sample.The probability density function (pdf) of a beta mixture model with two components (representing clean data and noise data respectively) on similarity C is defined as:

$$p(S|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} S^{\alpha-1}(1-S)^{\beta-1} \quad (5)$$

Among them, $\Gamma(\cdot)$ is the gamma function and S is the similarity,$\alpha,\beta > 0$.

Finally, the Expectation Maximization (EM) process was used to fit the BMM to the similarities of all samples in this study. After fitting the noise model, the probability that a sample belongs to a clean sample was obtained through the posterior probability:

$$w = p(m|S) = \frac{p(m)p(S|m)}{p(S)} \quad (6)$$

Among them, m = 1(0) represents the clean (noisy) class. $p(C|m)$ is defined as the posterior probability of the similarity S produced by component m.

C. Label Correction Module

The label correction module first obtains the label distribution of each sample through an auxiliary branch, and then uses class distribution mining to identify the potential invariants in the sample label distribution, thereby obtaining the category distribution of expressions. In the initial stage of training, the parameters of the auxiliary branch are unstable, which may lead to the inability to output a robust class distribution that accurately describes each facial expression category. To address this, a threshold t is introduced to

evaluate the stability of the class distribution in each period, which is expressed as follows:

$$d = \begin{cases} \frac{1}{=1} a & , \geq \\ & t \\ & , \\ & cl < \end{cases} \quad (7)$$

$$' = \begin{cases} = & \\ \frac{1-}{-1} & \neq \end{cases} \quad (8)$$

Among them, d^{c}_{class} is the class distribution of category c, d^{c}_{label} is the label distribution of the samples belonging to category c, $d^{c,y^{j}}_{t}$ represents the description degree of y^{j} in the threshold distribution corresponding to class c, and Nc is the number of samples belonging to class c.

Finally, the probability w that the samples obtained through the noise model are clean labels is used to correct the label distribution of each sample, as shown in the following formula:

$$f, = \cdot + \left(1 - \right) \cdot a, \quad (9)$$
$$cl,$$

D. Joint loss training

The auxiliary branch for generating the label distribution is trained and optimized using the CE loss, and the noise data during the training process is suppressed by the w obtained from the noise model to reduce the impact of noise on the model. For the target branch, the kld loss function and the modified label distribution are used for optimization training. In the first few training rounds, the focus is mainly on training the auxiliary branch to ensure that it can stably output the label distribution and category distribution. In the subsequent training rounds, the focus shifts to training the target branch and avoiding overfitting of the auxiliary branch. The loss function is as follows:

$$_1 = \cdot c \quad (10)$$

$$L = \mu_1 \cdot L_1 + \mu_2 L_{kld} \quad (11)$$

$$\mu_1 = \begin{cases} 1 & e < \beta \\ \exp\left(-\left(1 - \frac{\beta}{e}\right)^2\right) & e > \beta \end{cases} \quad (12)$$

$$\mu_2 = \begin{cases} \exp\left(-\left(1 - \frac{\beta}{e}\right)^2\right) & e < \beta \\ 1 & e > \beta \end{cases} \quad (13)$$

III. Experiment

A. Datasets

RAF-DB [8] (Real-World Affective Face Database) is a widely used real-world dataset in facial expression recognition research. This dataset was annotated by 40 well-trained human annotators and contains 15,339 facial images covering six basic expressions (happiness, surprise, sadness, anger, disgust, and fear) as well as a neutral expression. These images are sourced from real-life scenarios, aiming to provide researchers with a challenging environment to enhance the robustness of facial expression recognition algorithms.

In the experiments of this paper, 12,271 images were selected for training and 3,068 images for testing. This design enables a better assessment of the model's performance in real-world scenarios. Additionally, the overall sample accuracy of the test set is used as a crucial metric for evaluating the model's performance.

RAF-DB not only offers a rich collection of facial expression samples but also reflects the diversity of human emotions in daily life, providing valuable data support for the development and validation of facial expression recognition technologies.

B. Implementation Details

This paper trains the model under the environment of CUDA 11.3, PyTorch 1.12.0, torchvision 0.13.0 and Python 3.9, using NVIDIA RTX 3090. All images are further adjusted to a size of 100 × 100 before training. By default, the backbone network uses ResNet18, which is pre-trained on the MS-Celeb-1M dataset. For data augmentation, RandAugment [16], random horizontal flipping and random erasing [15] are adopted and performed online. The batch size for RAF-DB is set to 32. The threshold t is set to 0.7 by default, and the β value is set to 3 by default. The learning rate is initialized to 0.001. The Adam optimizer and ExponentialLR learning rate scheduler are used, with gamma set to 0.9 to decay the learning rate after each epoch. The training process ends at the 75th epoch.

C. Ablation study

To separately evaluate the roles of different modules, this paper conducted an ablation study on RAF-DB with a noise level of 30% using ResNet18 as the baseline model. To obtain more realistic performance, the average of the last five iteration results was used. The results are shown in Table I. From this, some conclusions can be drawn. The accuracy of the baseline model was improved by 8.07% by using the AEA module, which proves the effectiveness of the AEA module. Similarly, when the NP module was added to ResNet18+AEA, the accuracy was improved by 8.71%, which proves that the NP module can effectively reduce the impact of noise on the model and avoid overfitting to noisy data. In summary, the experimental results can well prove the effectiveness of the model.

Table I Ablation study

ResNet18	AEA	NP	30%RAF-DB
√			69.49
√	√		77.56
√	√	√	86.27

D. Performance Comparison

This paper quantitatively evaluates the improvement effect of the proposed model over other state-of-the-art noise-labeled facial expression recognition (FER) methods. The robustness of the model is explored by adding three noise ratios of 10%, 20%, and 30% on the RAF-DB dataset for a fair comparison. As shown in Table II, III, IV, the method proposed in this paper significantly outperforms other state-of-the-art FER noise label learning methods. For instance, the performance of the proposed method on RAF-DB is 6.91%, 8.52%, and 8.81% higher than that of thiocyanate respectively.

Table II Contrast experiment on 10% Noise RAF-DB

Methods	Noise	RAF-DB
SCN [12]	10	82.18
RUL [23]	10	86.22
EAC [21]	10	88.02
Ada-DF[26]	10	87.81
Our	**10**	**89.09**

Table III Contrast experiment on 20% Noise RAF-DB

Methods	Noise	RAF-DB
SCN [12]	20	80.10
RUL [23]	20	84.34
EAC [21]	20	86.05
Ada-DF[26]	20	86.67
Our	**20**	**88.62**

Table IV Contrast experiment on 30% Noise RAF-DB

Methods	Noise	RAF-DB
SCN [12]	30	77.46
RUL [23]	30	82.06
EAC [21]	30	84.42
Ada-DF[26]	30	84.38
Our	**30**	**86.27**

IV. Conclusion

In this paper, a dual-branch network framework for expression recognition is proposed. The framework of this paper includes an auxiliary branch and a target branch for facial expression classification. By extracting the label distribution, the category distribution of the entire dataset is mined. These category distributions and the label distribution of the data are adaptively fused through the noise probability obtained by the noise modeling module, taking advantage of both to obtain a more accurate fused distribution, thereby providing more accurate supervision for the training of the target branch. In addition, an average expression description module is introduced to suppress the problem of high inter-class similarity and intra-class difference, thereby improving the performance of the model. Through extensive experiments on the RAF-DB dataset, the effectiveness and robustness of the proposed method are demonstrated. In future work, we plan to explore more robust distribution generation by integrating more FER-related tasks, such as facial keypoint detection and facial action unit detection. Additionally, we aim to utilize multimodalities to extract more representative features, including 3D facial images, audio, and other relevant information sources. These efforts will contribute to the advancement of the FER field and expand the capabilities of the proposed method.

References

[1] Barsoum, E., Zhang, C., Ferrer, C.C., & Zhang, Z. (2016). Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the ACM International Conference on Multimodal Interaction* (pp. 279-283).

[2] Bazzo, J. J., & Lamar, M. V. (2004). Recognizing facial actions using Gabor wavelets with neutral face average difference. In *Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 505-510).

[3] Wang, Y., Ai, H., Wu, B., & Huang, C. (2004). Real time facial expression recognition with adaboost. In *Proceedings of the 17th International Conference on Pattern Recognition* (Vol. 3, pp. 926-929).

[4] Zhang, L., Chen, S., Wang, T., & Liu, Z. (2012). Automatic facial expression recognition based on hybrid features. *Energy Procedia*, 17, 1817-1823.

[5] 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)

[6] 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).

[7] Han, B., Yao, J., Niu, G., Zhou, M., Tsang, I., Zhang, Y., Sugiyama, M.: Masking: A new perspective of noisy supervision. In: NIPS (2018)

[8] 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).

[9] 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)

[10] 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.

[11] 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.

[12] Ruan, D., Yan, Y., Lai, S., Chai, Z., Shen, C., & Wang, H. (2021). Feature decomposition and reconstruction learning for effective facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7660-7669).

[13] 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).

[14] Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L.: Making deep neural networks robust to label noise: A loss correction approach. In: CVPR (2017)

[15] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 13001–13008.

[16] E. D. Cubuk, B. Zoph, J. Shlens, Q. 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, 2020, pp. 702–703.

[17] Ren, M., Zeng, W., Yang, B., Urtasun, R.: Learning to reweight examples for robust deep learning. In: ICML (2018)

[18] Wang, C., Wang, S., & Liang, G. (2019). Identity- and pose-robust facial expression recognition through adversarial feature learning. In *Proceedings of the ACM International Conference on Multimedia* (pp. 238-246).

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

[20] Yang, H., Ciftci, U.A., & Yin, L. (2018). Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2168-2177).

[21] 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.

[22] Zeng, J., Shan, S., & Chen, X. (2018). Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

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

[24] Xue, F., Wang, Q., & Guo, G. (2021). Transfer: Learning relation-aware facial expression representations with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3601-3610).

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

[26] 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.

[27] Ming, H., Lu, W., & Zhang, W. (2023). Soft label mining and average expression anchoring for facial expression recognition. In *Computer Vision - ACCV 2022: 16th Asian Conference on Computer Vision, Proceedings*, Lecture Notes in Computer Science, Vol. 13844 (pp. 728-744). Springer.

[28] Zhang, X., Lu, Y., Yan, H., Huang, J., Ji, Y., & Gu, Y. (2023). ReSup: Reliable label noise suppression for facial expression recognition. *arXiv preprint arXiv:2305.17895*.