EasyChair Preprint

№ 11287

# DacFER: Dual Attention Correction Learning for Efficient Facial Expression Recognition

Rui Sun, Zhaoli Zhang and Hai Liu



November 13, 2023

# DacFER: Dual Attention Correction Learning for Efficient Facial Expression Recognition

Rui Sun
*Faculty of Artificial Intell. in Education*
*Central China Normal University*
Wuhan, China
sunrui2021020556@mails.ccnu.edu.cn

ZhaoLi Zhang
*Faculty of Artificial Intell. in Education*
*Central China Normal University*
Wuhan, China
zl.zhang@ccnu.edu.cn

Hai Liu
*Faculty of Artificial Intell. in Education*
*Central China Normal University*
Wuhan, China
hailiu0204@ccnu.edu.cn

***Abstract*—Facial expression recognition is an important task in computer vision, the application of facial expression recognition in various fields continues to grow and its research is receiving increasing attention. However, sample noise and label noise are important challenges that cannot be ignored in facial expression recognition. The dual attention correction approach is proposed, which aims to raise the accuracy of local attention and focus on the importance of global attention. Specifically, the correction of local attention is reflected on the fact that the importance of each channel is increased through channel attention, as a result it suppresses useless features for the FER task, enhances useful features and prompts the classification loss function to obtain a more accurate basis of classification. The correction of global attention is reflected on the fact that more global information attracts attention with the help of spatial attentional shift consistency, therefore classification errors caused by local attentional "errors" are avoided. Under the influence of classification loss and spatial shift attention consistency loss, theDacFER method solves problems of input and label corruption and achieves recognition performance comparable to state-of-the-art methods of large-scale datasets RAF-DB and AffectNet in the wild. Our code will be made publicly available.**

***Keywords—Facial expression recognition, dual attention, noise label.***

## I. INTRODUCTION

Facial expression is one of the most visible and easily captured clues of human emotions [1]. Hence, facial expression recognition (FER) has a wide range of applications in education, healthcare, transportation, and human-computer interaction [2–4]. FER has a wide range of applications in the field of education. It can help teachers better understand and pay attention to students' learning status and needs, thereby improving the quality and effectiveness of teaching. Firstly, facial expression recognition can assist teachers in better classroom management. By observing students' facial expressions, teachers can assess their emotional and attentional states, and make timely adjustments to teaching strategies and classroom management methods. Secondly, facial expression recognition can help teachers personalize their teaching. By observing students' facial expressions and emotional states, teachers can assess their learning status and needs, and provide personalized instruction and guidance. Furthermore, facial expression recognition can be applied to online learning. In online learning, students often lack face-to-face interaction with teachers, making it difficult for teachers to assess their learning status and needs. In summary, the application of facial expression recognition in the field of education can help teachers better understand and pay attention to students' learning status and needs, thereby improving the quality and effectiveness of teaching. It has broad prospects for application in classroom management, personalized teaching, and online learning.

Recently, some excellent work has achieved superior performance in the FER benchmark datasets, nonetheless, uncertainty remains an important challenge for FER [5]. The uncertainty is primarily caused by noise, and the common FER dataset noise is divided into two categories: input noise and label noise. Specifically, input noise is the distortion at the image level. In Fig. 1, the difficulty in key feature extraction increases because facial texture and muscles have changed due to the age factor, therefore insufficient basis for deep learning classification leads to the ground-truth happy label being misclassified into the wrong sad label. Figure 1 is not identified as correspondingly correct labels owing to occlusion, lighting, makeup, blurring, posture and so on. Row 2 of Fig. 1 shows the noise label samples. The sample is labeled as happy, but its label is inclined to angry. The sample is labeled as surprise, but in fact its label should be neutral. The following samples also demonstrate the uncertainty of other labeled noise samples. The distribution of label noise of samples for deep learning capability is reflected in the fact that even through random label inputs, deep learning can completely remember noisy labels, degrading the generalization of the model [6]. The input noise increases feature extraction difficult because of the distortion of the image, so it cannot provide a basis for accurate classification.

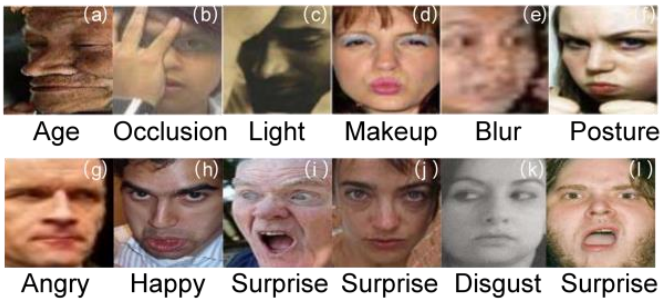


Fig. 1. Sample images selected on the RAF-DB dataset, the first row shows the input corrupted emoji images, the second row is the noise labeled facial images.

To mitigate the label noise and input noise issues, in-depth research has been done with excellent results. Firstly, the strategies to combat noisy labels can be divided into three categories: discarding, weighting, and modifying. Discarding is to remove the noisy labels from the training set. Weighting

refers to rank the reliability of noisy labels by noise-supervised methods. Modifying means the modification of possible noisy labels by regularized sorting. These strategies suppress noisy labels through the mixture of supervised classification loss and unsupervised consistency loss, focusing on the global significance of the samples. Secondly, as for input noise combat, more efficient feature extraction and key feature screening are required. At present, the feature extraction of CNN combined with attention mechanism is popular, which focuses on the local meaning of the image.

It is very important to pay attention on the global versus local meaning of image samples to mitigate FER uncertainty. To this end, we propose the Dual Attention Correction (DacFER) method with global and local attention for facial expression recognition. The major contributions of this paper can be summed as three aspects:

- Firstly, the channel attention mechanism is introduce into the DacFER method to enhance relevant features, suppress useless features, and enhance the feature extraction capability of the model against input corruption noise.
- Secondly, the spatial transformation attention consistency is proposed to correct the global attention of the samples against label noise.
- Thirdly, uniformly correcting local attention with global attention achieves excellent performance on large-scale benchmark datasets.

## II. Proposed Model

### A. Preliminary

Recently, facial expression recognition task has attacked many attention by the researcher. The research scenario changes from the laboratory to the wild, the mainframe network extends from CNN to Transformer. The loss function shifts from supervised classification loss function to a mixture of supervised classification loss and unsupervised consistency loss, the feature extraction moves from single high-level features to multi-scale semantics, and various outstanding work emerge in endless succession.

Currently, one of the most effective methods of local attention correction is the attention mechanism [8], including feature-related attentional mechanism, generic attentional mechanism, and query-related attentional mechanism [9]. To further improve the accuracy of face location, the authors introduced boundary prediction and CoordConv algorithm with boundary coordinates to deal with the imbalance between foreground and background pixels, and further apply spatial attention to realize the separation of face and background. In [10], Wang *et al.* proposed the RAN method, a new regional attention network, to adaptively capture the importance of facial region to gesture variable FER. Wang *et al.* [7] proposed a spatial attention method in order to capture key local similarities and adaptively weight different local patches. In [11], Farzaneh *et al.* proposed the DacFERL method [14], which has integrated an attention mechanism, and can estimate the attention weight value to represent the importance of features. This model can extract the intermediate space feature by CNN framework. The global attention is introduced to address the noisy labels problems. In fact, counteracting noisy labels benefits from multi-label learning technology. For example, Li *et al.* [12] proposed a spatial regularization network, which applies semi-supervised learning technology to generate the target probability according to the model output. Moreover, this method designed a regularization term to guide the model towards to these targets, implicitly preventing the memory of incorrect labels. Then, Li *et al.* proposed a contrast learning method for facial expression recognition task, which imposes smoothing constraints on adjacent samples to clear the noise, embeds the facial images into a low-dimensional subspace, and regularizes the geometric structure of the subspace by using robust contrast learning, including a mixture of unsupervised consistency and supervised classification loss. On the basis of the above work, the DacFER method is proposed for facial expression recognition.

### B. DacFER module

The overall architecture of our method is shown in Fig. 2, which mainly consists of a mainframe CNN, local attention correction and global attention correction. The mainframe CNN is leveraged to extract the feature maps. ResNet-50 is adopted here owing to its strong generalization ability and superior overall performance. For any given transformation, such as convolution, a feature mapping $U$ that maps the input to constructs an attention block for feature recalibration. The formula is denoted as,

$$U = Ftr(X). \tag{1}$$

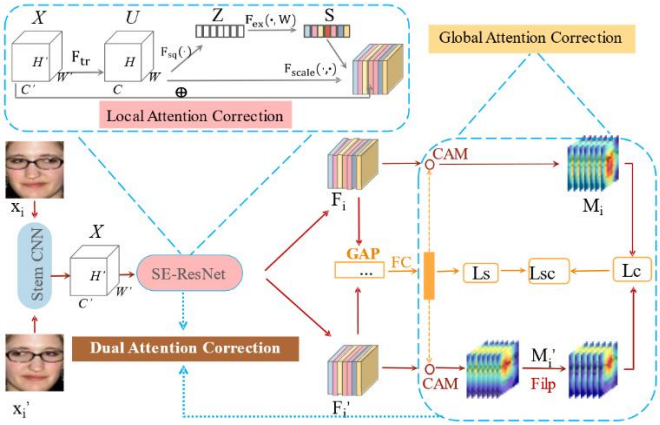


Fig. 2. Overall architecture of the dual attention correction model. First, the facial images are input to Stem CNN to extract feature maps. Second, through the channel attention mechanism, spatial channel dependencies are modeled and channel weights are reassigned to enhance useful channel features and suppress useless channel features. Third, while using softmax classification loss, the class activation mapping of the joint sample feature map is utilized to correct local attention and enhance global attention with the help of flipped sample feature map attention.

Class activation mapping (CAM) is a famous attention visualization method that allows us to visualize the class heat regions predicted on a specific image, highlighting the key regions of interest for the CNN model. Dimensionality reduction, parameter reduction, and assigning classification meaning to channels are achieved by global level pooling of the feature map *Fi*. Full connectivity is utilized as a classifier for classification. Note that the graph is a weighted sum of the feature maps of the third convolutional layer and the weights of the fully connected layer, and the weighting is done by a product operation with the following equation:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \ , \quad (2)$$

where the weights of FC layers is denoted by $W \in R^{K \times C}$, and $K$ means the number of classes. Attention graph calculation means $M_j$, in which $M_j(h, w)$ is the attention value of location $(h, w)$ for class index $j$ and the weighted sum of feature maps over different channels. In this model, the famous CAM is employed to calculate attention mappings from the input image to show the features that the model pays attention to.

DacFER method applies *Softmax* loss for basic classification and corrects local attention. $M$ denotes the number of samples in a batch, $y_i$ is the true label, $K$ denotes the total number of categories, and it extracts the d-dimensional features $x_i \in R^{d \times 1 \times 1}$,$w \in R^{d \times K}$,$b \in R^{K \times 1}$, in which $w$, $b$, are the weights and biases of the fully connected layer, respectively. As for $L_s$, the probability distribution P is calculated on the category where the *Softmax* function is located, and finally, the cross-entropy loss is calculated as the difference between the predicted and true label values, and the final *Softmax* loss is derived as follows,

$$\begin{aligned} L_s &= -\frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{K} y_i log\, p(y = j \mid x_i) \\ &= -\frac{1}{M} \sum_{i=1}^{M} log \frac{e^{w_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{K} e^{w_j^T x_i + b_j}} \end{aligned} \quad . \quad (3)$$

The correction of global attention is supervised by metric learning, in which $M_{ij}$ denotes the weighted feature map, and $(M^{'})_{ij}$ denotes the weighted feature map of the flipped samples,

$$L_c = \frac{1}{MKHW} \sum_{i=1}^{M} \sum_{j=1}^{K} \| M_{ij} - Flip(M^{'})_{ij} \|_2, \quad (4)$$

where $L_s$and $L_c$ jointly supervise attention correction training processes. Both local attention and global attention rely on loss functions for optimization iterations,

$$L_{total} = L_s + \lambda L_c, \quad (5)$$

where $\lambda$ represents the coefficient of consistency loss, which is utilized to control the contribution of consistency loss to the network.

## III. Experimental Results and Discussion

### A. Implementation Detail

The SE-ResNet-50 is introduced as the backbone in our experiment. The pre-training model was trained on the MS-Celeb-1M dataset and then fine-tuned on the VGGFace2 dataset. In our experiment, the standard Adaptive Moment Estimation (Adam) optimizer is leveraged for weight attenuation of 1×10−4. All the input images are enhanced in real time by randomly erasing an area of the image to zero. The input facial images are cropped the central region in the testing stage. A crop of size 224×224 was extracted from a 256×256 input facial images. The proposed DacFER is trained 60 rounds on RAF-DB with an initial learning rate of 0.001 and attenuation of 0.9 times per round. Alternatively, we trained ResNet on AffectNet for 20 cycles with an initial learning rate of 0.0001 and a 5-fold decay in each five cycles. The batch size is set as 64 for both datasets. The parameter $\lambda$ is equal to 1 in this article. Under our particular trunk architecture setup, the depth features are 2048 dimensional, and the last convolution feature map x×1 has a size of 2084×7×7. In this work, the pytorch deep learning framework is utilized to train the proposed DacFRE model on a NVIDIA RTXA4000 GPU with 16GBV-RAM. Conducting experiments on RAF-DB, AffectNet, and FERPlus datasets.

### B. Experimental results

In Table 1, it shows the recognition accuracy of the DacFER method on the RAF-DB datasets. Since RAF-DB is an unbalanced dataset, we also report the average recognition accuracy. Experiments on the RAF-DB dataset demonstrate that our method outperforms the current state-of-the-art methods, achieving an overall recognition accuracy of 90.61% and an average accuracy of 83.14%. In Table 2, it shows the recognition accuracy of the DacFER method on the AffectNet datasets. Experimental results on AffectNet dataset demonstrate that the proposed method can achieve 65.46% recognition accuracy, which is better than other comparing FER method.

TABLE 1. STANDARD AND AVERAGE ACCURACIES WERE COMPARED ON THE RAF-DB DATASET.

| Methods | Acc. (%) | Avg.Acc (%) |
|---|---|---|
| FSN | 81.10 | - |
| pACNN [15] | 83.27 | - |
| DLP-CNN | 84.13 | 74.20 |
| ALT | 84.50 | 76.50 |
| gACNN | 85.07 | - |
| IPA2LT | 86.77 | - |
| RAN [10] | 86.90 | - |
| SCN | 87.03 | - |
| DACL [11] | 87.78 | 80.44 |
| KTN [16] | 88.07 | - |
| DMUE [1] | 88.76 | - |
| RUL [5] | 88.98 | - |
| EAC [13] | 90.35 | - |
| Proposed DacFER | **90.61** | **83.14** |

TABLE 2. EXPRESSION RECOGNITION ACCURACY OF VARIOUS METHODS WAS COMPARED ON AFFECTNET DATASET

| Methods | Acc. (%) |
|---|---|
| pACNN [15] | 55.33 |
| IPA2LT | 57.31 |
| IPFR [17] | 57.40 |
| separate loss [18] | 58.89 |
| DDA loss | 62.34 |
| RAN [10] | 59.50 |
| SCN | 60.23 |
| DACL [11] | 65.20 |
| KTN [16] | 63.97 |
| DMUE [1] | 62.84 |
| RUL [5] | 61.43 |
| EAC [13] | 65.32 |
| Proposed DacFER | **65.46** |

To achieve a more detailed understanding of the experimental results, we visualize the performance of the DacFER method for recognizing various facial expressions using a confusion matrix.

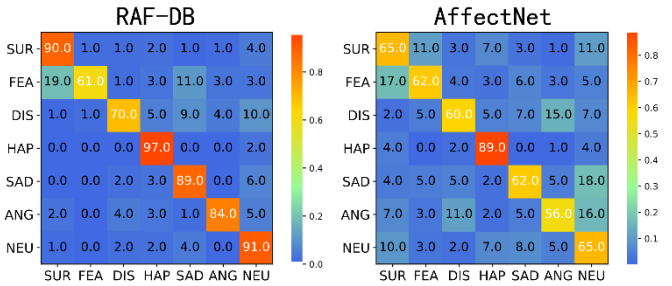


Fig. 3. The confusion matrix: visualizing the recognition performance of the DacFER method on RAF-DB and AffectNet for different facial expressions.

In Fig. 3, it can be found that the proposed DacFER method on the RAF-DB can achieve an accuracy of over 80% for the recognition of emotions other than disgust and fear. Among them, the accuracy for recognizing happiness is the highest, reaching 97%. However, the recognition accuracy for disgust and fear is relatively low. This imbalance in recognition accuracy is caused by the uneven distribution of sample quantities. On the AffectNet, happy achieved the highest recognition accuracy, reaching 89%, while the recognition rates for other types of expressions were around 60%. This indicates that in a balanced dataset, the distribution of recognition accuracy among different categories is relatively balanced. However, the long-tail distribution problem in facial expression recognition tasks still needs to be addressed. It can be conclude that the proposed DacFER method has shown excellent performance in various expression recognition tasks on the two datasets, but the negative impact of long-tail distribution samples on the model has not been eliminated.

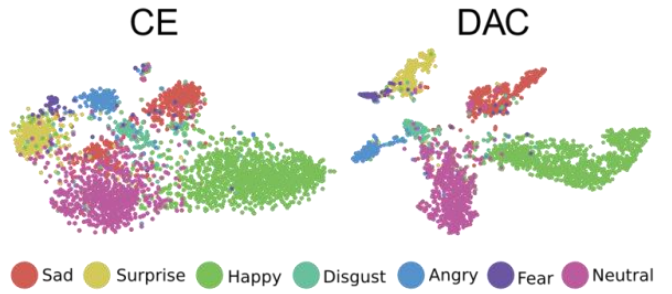


Fig. 4. Visualization of deep learning features, using the t-SNE method is shown. Left, right are the distribution of deep learning features under baseline method, DacFER method, respectively.

Figure 4 shows the feature vectors after t-SNE visualization, and the comparison shows that DacFER increases the inter-class separability and reduces the intra-class variability compared with the CE method. The DacFER method has superior classification performance.

### C. *Ablation study*

To further elucidate the effectiveness and work theory of the DacFER method, we conducted the following ablation experiments: 1) Compare the recognition accuracy rate of the baseline method with the LAC and GAC methods, and deconstruct the contribution of each part of the DacFER method to the model performance improvement; 2) Visualize the attentional focus regions of the baseline, LAC, GAC, and DacFER methods on the samples by the GAM method; 3) Compare the DacFER method with the baseline method on RAF-DB dataset and AffectNet dataset to show more experimental details.

Table 3 shows the recognition accuracy ablation experiments for the RAF-DB and AffectNet datasets. On the AffectNet, the recognition accuracy of the baseline method is 62.09%, the LAC method with the addition of channel attention improves 1.23% on this basis, the GAC with the addition of flip attention correction improves 2.14%, and the DacFER method improves 3.37%, which shows that both the LAC method and the GAC method can effectively improve the recognition accuracy of the model. The ablation experiments on the RAF-DB also proved the above conclusions.

TABLE 3. EVALUATE THE ABILITY OF CNN, LAC, GAC TO CONTRIBUTE TO DACFER METHOD ON RAF-DB,AFFECTNET DATASET.

| CNN | LAC | GAC | RAF-DB (%) | AffectNet (%) |
|---|---|---|---|---|
| ✓ | - | - | 88.66 | 62.09 |
| ✓ | ✓ | - | 88.98 | 63.32 |
| ✓ | ✓ | ✓ | 90.61 | 65.46 |

Furthermore, the GAM method is utilized to visualize the attentional focus regions of the various methods on the samples, as shown in Fig. 5. The baseline method focuses on some regions of the face but ignores the global significance of the face. The classification will easily fail if the local attention is wrong. For this reason, two strategies can improve this deficiency. One is to ensure that the local attention make as few errors as possible, and the other is to pay attention to the global meaning of the face. To this end, LAC enhances the feature extraction ability of the model by suppressing useless features through the channel attention mechanism to correct local attention and reduce the possibility of local attention errors. On the other hand, global attention cannot be neglected, and the GAC method with the help of flipped attention correction focuses on more emotional details of the face and improves the error tolerance of the baseline method. With the combined effect of LAC and GAC methods, the DacFER method suppresses the uncertainty caused by light, facial condition, age, occlusion, blurring and so on, which improves the robustness and generalization ability of the model.

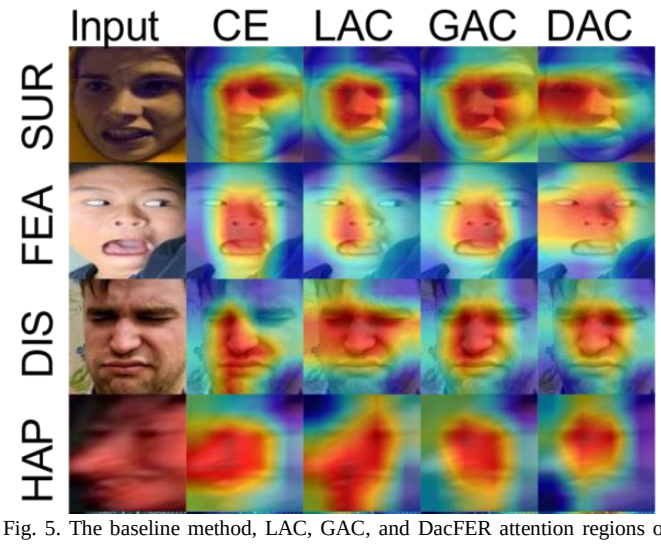


Fig. 5. The baseline method, LAC, GAC, and DacFER attention regions on noisy samples were compared.

To demonstrate the effectiveness of the DacFER method against noisy labels, we trained the RAF-DB dataset with 10%, 20%, and 30% noisy labels respectively. In Table 4, compared with the current state-of-the-art noisy label countermeasures, the DacFER method improved 0.54%, 1.27%, 0.52% of recognition accuracy, which proves that the DacFER method improves the ability of the model to fight against noisy labels.

TABLE 4. COMPARE THE RECOGNITION ACCURACY OF VARIOUS NOISE COUNTERMEASURES ON THE RAF-DB DATASET FOR NOISY LABELS

| Methods | Noise (%) | | |
|---|---|---|---|
| | 10 | 20 | 30 |
| Baseline | 81.01 | 77.98 | 75.50 |
| SCN(CVPR20) | 82.15 | 79.79 | 77.45 |
| RUL(NeurIPS21) | 86.17 | 84.32 | 82.06 |
| EAC (EECV22) | 88.02 | 86.05 | 84.42 |
| DacFER(Ours) | **88.56** | **87.32** | **84.94** |

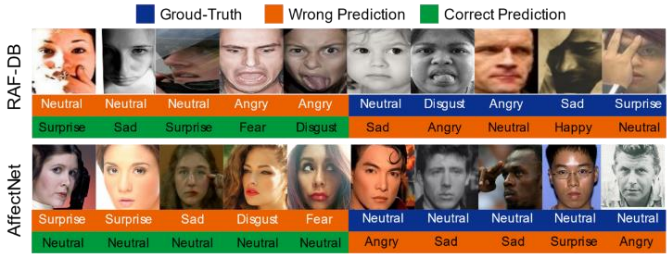


Fig. 6. Training model with DAC method, some correctly classified and misclassified sample labels from RAF-DB, AffectNet.

As shown in Fig. 6., by comparing the corrected mislabeled samples on the RAF-DB and AffectNet datasets, we found that the DacFER method has suppressive effect on uncertainties caused by facial condition, light, and occlusion, but light, age, blurring, and occlusion are still the main challenges faced by the FER task, especially when a sample has multiple uncertainties at the same time, the difficulty of accurately identifying the sample is greatly increased.

## IV. CONCLUSION

In this work, we proposed an efficient facial expression recognition network with the channel attention mechanism and spatial attention mechanism. Specially, the channel attention mechanism is introduced to address the input corruption issue, which counteracts noise problem in the low-quality images. Secondly, the spatial attention is proposed to focus on the importance of global attention. These two approaches together suppress the uncertainty problem in FER. Experimental result demonstrate that the proposed method has achieved good performance on FER noisy datasets and it can meet the challenges of FER noisy datasets in special cases. In the future, we will examine the fast speed calculation on the proposed DacFER model, can works on the video scenarios.

## REFERENCES


[1] J. Shen, Y. Hui, H. Si, J. Wan, H. Sheng, T. Men, Dive into ambiguity: Latent distribution minings and pairwise uncertainties estimation for facal expression recognitions, in: Proceeding of the IEEE/CVF Conferences on Compute Vison and Patten Recogntion (CVPR), 2022: pp. 6249–6258.

[2] X. Fang, Z. Den, K. Wan, X. Pen, Z. Qiang, Learning discriminative representations for facial expressions recognition from unceretainties, in: Imaging Processings (ICIP)for IEEE Internatonal Confrence, 2022: pp. 905–910.

[3] D. Gera, S. Balasubramanian, Noisy annotations robust consensual collaborative affect expression recognition, in: Proceedings of the IEEE/CVF Computer Vision for International Conferences (ICCVs) Workshop, 2022: pp. 3586–3593.

[4] F. Zhan, M. Xiu, C. Xiu, Weakly-supervised facial expression recognition in the wild with noisy data, IEEE Transactions on Multimedia, 2022: pp. 1800–1814..

[5] Y. Zang, C. Wang, W. Deng, Relative uncertainty learning for facial expression recognition, Advances in Neural Information Processing Systems, Curran Associates, Inc., 2021: pp. 17616–17627.

[6] L. Jang et al., MentorNet: Very deep neural network for learnig data-driven curriculum on corrupted labels, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, PMLR, 2018: pp. 2304–2314.

[7] Q. Wan et al., LS-CNN: Multiples scale for face recognition by Characterizings loecal patchs, IEEE Trans. on Information Forenscs and Securiety, 2023: pp. 1641–1654.

[8] F. Xue, *et al*., Robust facial expresion recognitions by Vison transformers with attentive poolings, IEEE Transa. on Affectve Computng, 2022: pp. 12–25.

[9] G. Brauwee et al. , Attention mechanims in deep learnings by a genaral survey, IEEE Transa. on Knowladge and Date Enginering, 2023: pp. 3278–3297.

[10] K. Wag et al. , Region attetion networks for pose and occlusion robust facial expresion recognitions, IEEE Trans. on Imaging Processings. 2021: pp. 4057–4069.

[11] A.H. Fareznaeh et al. , The wild facial expression recognition via deep attentive ceter loss, in: Proceeding of the IEEE/CVF Winte Confernce on Computer Vision Applicaions, 2023: pp. 2403–2413.

[12] S. Liu et al. , Prevents memorization of noisy labels by Early-learning regularization, in: H. Larochele, M. Rarzato, R. Hadsell, M.F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc., 2020: pp. 20331–20342.

[13] N. Zhang, *et al.* , Learning from all: noisy label facial expression recognitions by erasing attention consistency, in: Computer Vision –17th European Conference, Tel Aviav, Iarael, October 24–28, 2022, Proceedings, Heidlberge, 2022: pp. 418–434.

[14] M. Ren et al., Robust deep learning by reweight examples, in: J. Dy, A. Krause (Eds.), Proceedings of the 36th International Conference on Machine Learning, PMLR, 2019: pp. 4331–4343.

[15] Y. Liu *et al.* , Occlusion-awarre facial expression recogition by patch-gated CNN, in: 2019 25th Internatinral Confereces on Patern Receogniton (ICPR), 2019: pp. 2208–2215.

[16] H. Liu *et al*., The C-F label and distillation of adaptiely learning facial expression representation, IEEE Trans, on Imege Processing. 30 (2022) : pp. 2017–2029.

[17] C. Wan *et al* , Adversarial feature learnng in identities-and pose-robust facial expressions recognitions, in: Proceedings of the 28th ACM International Conference on Multimedias, Assocation for Computing Machinery, New Yorks, USA, 2019: pp. 237–249.

[18] Y. Liu *et al*., Basic and compouned faciael expresion recogniton in the wild Seprarate loss, Proceding of the Elevenths Asisans Confernces on Machinse Leanineg, PMLRs, 2020: pp. 898–912.