# Facial Expression Recognition by Regional Attention and Multi-task Learning

Longlei Cui, Ying Tian

**Abstract — Facial expression recognition(FER) is an significant branch of artificial intelligence and plays an indispensable role in optimizing and improving the experience of HCI. Although much progress has been made in this field, there is a need to optimize the accuracy rate of facial expression recognition and the network's generalization ability. This Paper brings forward a new end-to-end region attention and multitask learning network (RAMN) for FER. The method learns the importance of the facial features and combines different numbers of regional features obtained from the neural network for FER. A region loss function (R-loss) and collaborative loss function (C-loss) are used to jointly optimize RAMN. The R-loss can learn the most discriminative region in the facial images, and the C-loss substantially reduces the intra-class differences and increases the intra-class differences in the facial features. Finally, experiments on two uncontrolled datasets (FER2013, RAF-DB) and two controlled datasets (CK +, Oulu-CASIA) to verify the superior performance ability of the proposed method. A contrast with existing methods demonstrates the superiority of the RAMN method for FER.**

**Index Terms—Deep Convolutional Neural Network(DCNN), Facial Expression Recognition, Region Attention, Multi-task Learning Network**

## I. Introduction

Facial Expression Recognition refers to the ability to extract and recognize facial expression information. Humans can express and recognize emotions, and facial expressions communicate this information. Scientists have demonstrated that human beings have internal and external mechanisms, including measurable and objective emotional responses (e.g., neutral, sadness, surprise, fear, and disgust). The ability to express these mental states through facial expressions is an innate human ability that is essential in our everyday communication and social interactions. Scientists in the fields of computer vision and human-computer interaction neuroscience [1], [2] have done a lot of research on facial expression. DCNNs have recently shown excellent score for various image multi-classification tasks [3], [4]. By carefully designing the convolution, pooling, and layered architecture, local and global features were used to derive rich visual information; thus, DCNN has been recognized as a very effective tool for FER [5]. The challenges of facial recognition research, such as EmotiW series and Kaggle's FER challenges, show that deep learning has become a trend to solve this problem. Most previous FER studies processed the entire image of the face without considering the exact position of the face in the image [6], resulting in unnecessary computational costs. Human FER in the laboratory environment has achieved good results, but the FER performance under real conditions requires improvement. The attention mechanism has been extensively analyzed recently [7]-[9], providing a new research direction for existing neural network models [10], [11]. Neuroscientists have also investigated the attention mechanism [12] since it is believed to be crucial to visual perception and cognition and can affect the information conveyed by the visual cortex. In addition, it has been shown that in multitask learning, the use of shared features in the training of multiple related tasks can enhance the generalization ability of the network [13].

This article mainly includes the following three aspects:

(1) We propose an region attention multitask learning network (RAMN) for FER.

(2) A collaborative loss function (C-Loss) and region loss function (R-Loss) are incorporated into the RAMN to enhance the network. The R-loss can learn the position with the most discriminative area in the facial image and weigh the important position of human face. The C-loss substantially reduces the intra-class differences and increases the interclass differences in the facial features to up-grading the accuracy rate of FER.

(3) Experiments conducted on two uncontrolled datasets (FER2013, RAF-DB) and two controlled datasets (CK+, Oulu-CASIA) demonstrate the superiority of the RAMN for FER. The test accuracy rates are 75.14% on FER2013, 97.85% on CK+, 86.87% on RAF-DB, and 87.79% on Oulu-CASIA.

## II. Related Works

Typically, FER includes three main steps: facial image preprocessing, image feature extraction, and expression classification. Various face detection algorithms have been used for preprocessing to detect faces and perform face alignment operations, such as the multi-task cascaded convolutional network (MTCNN) [14] and Dlib [15]. Different feature extraction methods were developed to get the facial geometry and features of facial expressions. These methods can be divided into those using basic features and engineering features. The latter include methods

Manuscript received December 17, 2020; revised Jun 2, 2021.

This work was funded by the foundation of Liaoning Educational committee under the Grant No.2019LNJC03.

Longlei Cui is a Master Student of School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan, China (e-mail: asliyou@126.com).

Ying Tian, the corresponding author, is a Professor of School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan, China.(phone: +8613898015263; fax: 0412-5929818; e-mail: astianying@126.com).

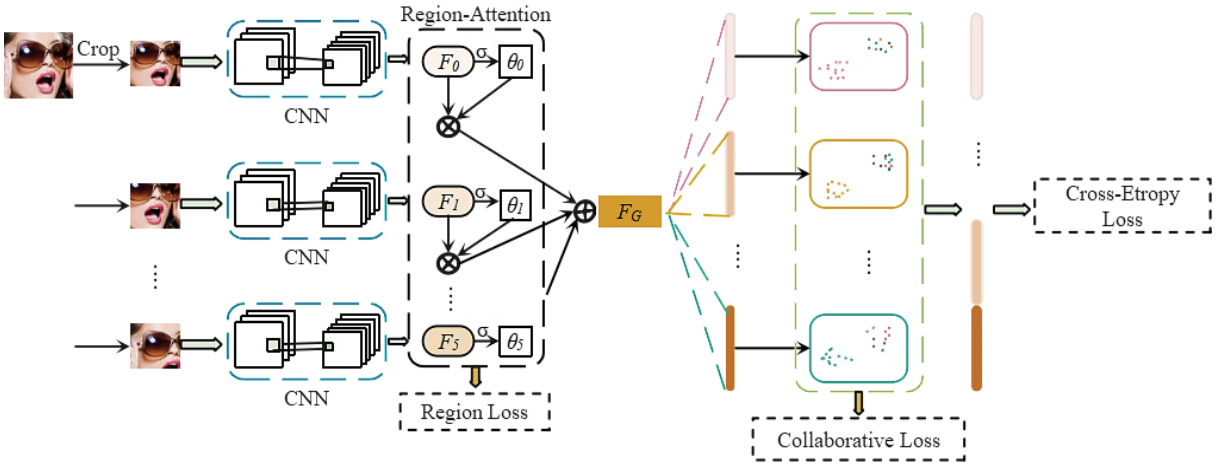



Fig. 1. System architecture diagram of the RAMN.

based on geometry-based global features, texture-based local features, and mixed features. Feature extraction methods based on geometric features use landmark around the nose, mouth, and eyes. Methods to extract texture-based local features include scale-invariant feature transform (SIFT) [16], and Gabor wavelet coefficients [17], local binary pattern (LBP) [18] histogram. Having two or more engineering characteristics is referred to as mixed feature extraction, which is used to improve the characterization results. Regarding deep learning, Fasel [19] found that shallow CNNs were robust for different scales. Liu et al. [20] proposed a CNN structure based on a facial motion unit for FER. Subsequently, the features were input into a SVM [21], and a softmax layer, logistic regression, and other supervised classifiers were used to classify the expressions.

Many recent studies [22]-[27] used FER datasets to pre-train the neural network and optimized it to avoid over-fitting due to small sample sizes. Levi and Hassner [25] used the CASIA-Web Face FER dataset to pre-train at least four different GoogleNet [28] and VGGNet [29] networks. Zhao et al. [26] proposed a peak gradient suppression (PGS) programme, and pre-trained the model on the CASIA-Web Face dataset. Meng et al. evaluated different FER model architectures and used FER datasets. Ding et al. [27] proposed a FaceNet2ExpNet framework, which combines face recognition tasks and FER training. Albanie et al. used the VGG Face model and fine-tuned FER Plus with soft probability. Liu et al. [30] used a spatiotemporal manifold feature based on a middle representation to extract features from a facial image sequence. The region attention network (RAN) [31], which uses the self-attention module and relationship attention module for facial recognition, is probably the most similar to the model proposed in this paper. The difference is that our model uses a region attention module and multitask learning module.

## III. Proposed Methods

In this section, the proposed RAMN is introduced, describes the modules in detail, and presents the method to clip the facial area from the images.

### *A. Overview of the proposed methodology*

Due to the importance of local information in FER and effective measures such as mutual learning between multiple tasks, the generalization ability of the network requires improvement. In this paper, RAMN is put forward to surmount the shortcomings of the CNN model based on global features and single-task learning. The proposed network adaptively captures the importance of the local information in the face, providing a reasonable trade-off between local and global features and the intra-class and interclass differences in facial expressions. The system detailed structure diagram of the RAMN is shown in Fig. 1. The system consists of three parts: the feature extraction module, region attention module, and expression classification module. After face alignment, the face image is subdivided into several areas (including the original image), which are input into the backbone CNN model(VGG16) for feature extraction. The region attention module uses the fully-connected (FC) layer and a sigmoid function to assign attention weights based on each clipped image. The R-loss constrains the weight of the region attention module to determine the importance of the clipped images. The weights of the region attention module and the features extracted from the CNN are weighted to each clipping region to become a shared feature($F_G$ in Fig. 1). Multiple parallel FC layers are used to analyze the features of multiple categories. Each FC layer corresponds to an expression of the learning tasks and generates the homologous expression features. The expressions of the feature vector are combined, and the standard cross-entropy loss function is used for predicting the expression categories. All feature learning tasks for a single feature are performed simultaneously using the C-loss.

### *B. RAMN*

Figure. 1 is the system structure diagram of RAMN, which is divided into two stages. In the first stage, the FC layer and a sigmoid function are used to calculate each region's features, and the shared feature $F_G$ is obtained. In the second stage, multitask learning is adopted. Multiple FC layers are connected in parallel, and each FC layer corresponds to an expression learning task to perform the expression classification.

#### *1) Region Attention module*

CNN training is performed on the clipped image to obtain a feature representation. According to the regional features, t

he FC layer and a sigmoid function are used in the region att ention module to estimate the attention weights of the clippe d images. In this phase, the regional features and attention we ights are combined into the shared feature $F_G$:

$$F_G = \frac{1}{\sum_{k=0}^{m} \theta_k} \sum_{k=0}^{m} \theta_k F_k \tag{1}$$

The attention weight of the $k$-th region is defined as $\theta_k$, and $F_k$ is the facial feature representation of the $k$-th region. $F_G$ is the shared feature used for facial expression classification.

*2)Region Loss*

Since it is assumed that the cropped face image has more discriminant features than the original image, a constraint is imposed on the attention weight of the region attention module, i.e., the R-Loss. It is defined as:

$$L_R = \max\{\beta - (\theta_{\max} - \theta_0), 0\} \tag{2}$$

where $\beta$ is a parameter in a certain range, $\theta_{max}$ is the maximum weight of the facial region, and $\theta_0$ is the weight of the cropped face image.

*3)Collaborative Loss*

The CNN network obtains the category of the facial expression of each input image. For the $i$-th facial expression, the feature information obtained from the $j$-th sample is input into the CNN network and is denoted as $x_j^i$. If there are two points $d_0^i$ *and* $d_1^i$ in the feature space, $d_0^i$ represents the feature center that does not belong to the $i$-th class of the facial expression, and $d_1^i$ represents the feature center of the $i$-th class of facial expression. The distance between the two feature centers is:

$$l_{pos,j}^i = \left\| x_j^i - d_{y_j^i}^i \right\|_2^2 \tag{3}$$

$$l_{neg,j}^i = \left\| x_j^i - d_{1\text{-}y_j^i}^i \right\|_2^2 \tag{4}$$

Where $y_j^i \in \{0,1\}, y_j^i = 1$ illustrates that the $i$-th facial expression contains $x_j^i$, $y_j^i = 0$ explains that the $i$-th facial expression does not contain $x_j^i$, $l_{pos,j}^i$ is the positive square of the Euclidean distance between a facial image feature and the center of this class, $l_{neg,j}^i$ is the negative square of the Euclidean distance between a facial image feature and the center of this another class, and $\|.\|_2$ represents the Euclidean distance. The intra-class loss of the facial expressions of the $i$-th type is represented as $l_{pos,j}^i$ or $L_{M,j}^i$. $L_{M,j}^i$ describes the intra-class differences in the facial expression sample features but not the inter-class differences.

The following constraint is added to the sample facial expressions to improve the discriminability of the facial features and ensure that intra-class cohesion and interclass discernibility of the sample facial expressions are considered:

$$l_{pos,j}^i + \omega < l_{neg,j}^i \tag{5}$$

This constraint ensures that the negative distance $l_{neg,j}^i$ of the facial expression sample features is greater than the positive distance $l_{pos,j}^i$ by a specific distance $\omega$. Thus, the feature points of one type of expression are far from the feature center of the other types in the feature space, increasing the feature distance between the expressions. According to the above constraints, we define the interclass loss of the samples of the $i$-th facial expression as follows:

$$L_{N,j}^i = \max\{\omega + l_{pos,j}^i - l_{neg,j}^i, 0\} \tag{6}$$

This loss penalizes feature points with insufficient negative distances, where the distance $\omega$ is a non-negative number. $L_{N,j}^i$ and $L_{M,j}^i$ constitute the loss of a given facial expression sample in the acquiring process of the network.

During training, the R-Loss, C-Loss, and cross-entropy Loss are applied to the combination to facilitate the network's traing of the facial expression features.

*C. Area cropping methods*

Cropping out different regions from the face images is a essential step in facial expression detection. The cropping strategy has a substantial influence on the classification results. If the image is cropped into too many parts, redundant image information is produced, affecting the network training efficiency. If there are too few parts, the network cannot learn useful features. We introduced the parameter $\beta$ to enable the network to learn useful features from relatively few facial images. In addition, cropping the image also increases the sample size. Therefore, we cropped the original image into 5, 7, and 9 sub-images to evaluate the three cropping methods. Fig. 2 shows the proposed cropping method.

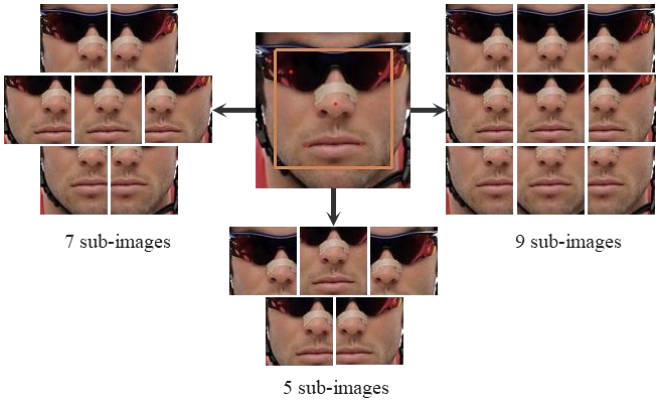



Fig. 2. Facial area cropping method.

**5 sub-images:** The four quadrants of the image (upper right, upper left, lower right, and lower left) were used as a reference, and the image was clipped at a 3/4 scale. The center of the image (0.8×0.8) was clipped again to generate 5 sub-images.

**7 sub-images:** After clipping the image into 5 sub-images, the center box moves from left to right. The left and right sides of the center box are overlapped with the left and right sides of the image to generate 7 sub-images.

**9 sub-images:** Based on 7 sub-images, cut the center box up and down once each, generating 9 sub-images.

## IV Experimental Results And Analysis

The datasets are introduced, In this section, and the experimental results obtained from RAMN and other models are compared.

*A.Datasets*

**FER2013**. The FER2013 dataset is large with 35,886 images, of which the training images is the largest. All images are aligned and have a size of 48×48 pixels in the dataset. FER2013 includes 7 emotions, i.e., happy, neutral, surprised, disgusted, angry, sad, and fearful.

**CK+**. The CK+ dataset is divided into 7 expressions. All images were obtained in a laboratory environment. The expression of each volunteer changes from a neutral expression to the strongest state of an expression. The last three frames were extracted from the sequence of each subject in the CK+ dataset, containing a total of 981 facial expressions. A 10-fold cross-validation was used on this dataset.

**Oulu-CASIA**. The Oulu-CASIA dataset is similar to the CK+ dataset and has six expressions. Each video sequence begins with a neutral face and ends with a peak facial expression. Similar to the experimental setup of CK+, the last three frames of the images were extracted from each subject sequence in the Oulu-CASIA dataset; 240 images were collected for each category, with 1,440 emotions. A 10-fold cross-validation was performed on this dataset.

**RAF-DB**. This dataset contains various facial expression images, such as occlusions and large posture; thus, this dataset is more challenging. It has an unequal number of expression categories in the basic expressions. For example, there are only 355 images with a fearful expression and 5,957 images depicting happiness. We selected only images with basic emotions: Among them, 12,271 images were used for training data and 3,068 for validation data.

*B. The experimental details*

In the following experiments, we used the face alignment method of the MTCNN to crop and align the face. The cropped image of the face was sized to 224×224 pixels. PyTorch was used to implement the proposed model. For the backbone CNN, the VGG16 network model trained on the CASIA-Web Face was used, and fine-tuning was conducted on different datasets. In all datasets, the learning rate was initialized to 0.01, and after 30 epochs, the learning rate was adjusted to 0.9 for every 5 epochs that were removed; the training was stopped after 60 epochs.

*C. Area cropping evaluation*

We used β=0.02 and the default parameters on different data sets to evaluate the model performance for the 5, 7 and 9 sub-images. The experimental data are represented in Fig. 3(a). Among the three cropping modes, the model performance was better for the 7 and 9 sub-images than the 5 sub-images. However, there was only a small performance improvement between the 7 and 9 sub-images. This result indicated that only a negligible amount of additional information was extracted from the 7, 9 than the 5 sub-images.

Therefore, we evaluated the model performance for 4 sub-images (upper right, upper left, lower right, and lower left) and 10 sub-images (the original 9 sub-images and a randomly cropped image). The model performance was lower for the 4 sub-images than the 10 sub-images. In addition, the model performance was about 0.1% higher for the 10 sub-images than the 5 sub-images. The reason is that

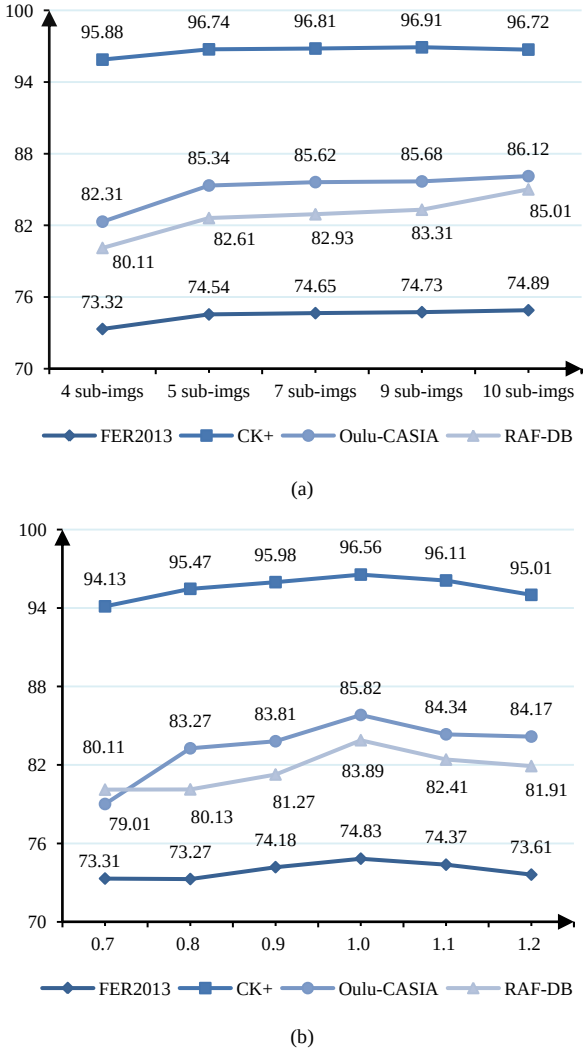


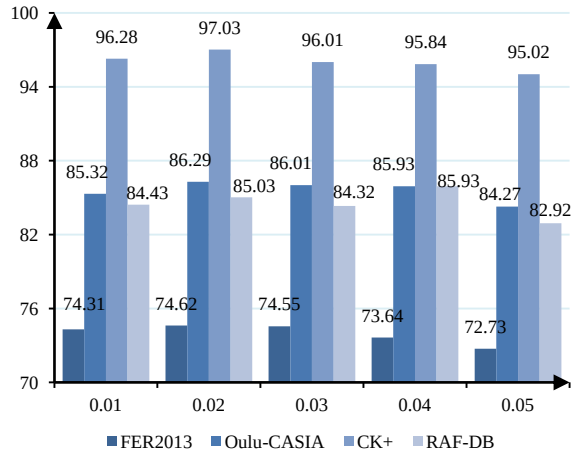

Fig.3. (a) Comparison of the network performance for the five cropping methods on different datasets; (b) Comparison of the network performance using the fine-grained features on different datasets.

the 4 sub-images do not contain sufficient information. For the 10 sub-images, the random cropping increases the network performance. However, a large sample size significantly increases the calculation amount and the use of GPU resources. The experimental results showed that the initial assumption was correct, i.e., the parameter $\beta$ ensured that the network learned important facial features from relatively few images. It was also observed that some cropped images were more discriminative than the original images. In the following experiments, 5 sub-images were used. Subsequently, we analyzed the effects of fine-grained cropping on the network performance for the same value of parameter $\beta$. Six fine-grained cropping tests were carried out on different data sets, and the image was cropped 0.7, 0.8, 0.9, 1.0, 1.1, and 1.2 times. Fig. 3(b) illustrates the results of this experiment. If the benchmark scale increased or decreased slightly, the model performance was reduced. The model performance was reduced compared to the benchmark scale of 1.1 times. The likely reason is that the cropped face area was too large, and the network could not learn the fine-grained features of the face.

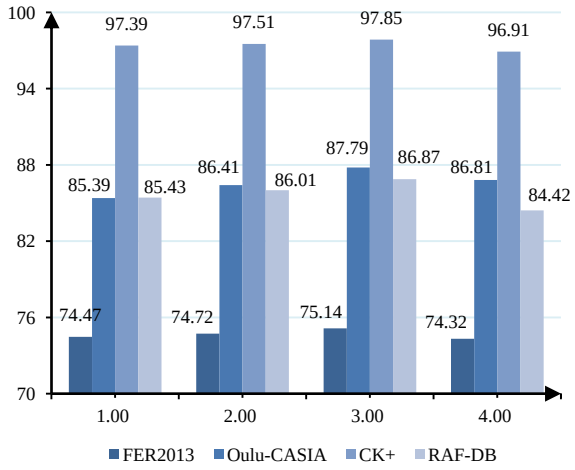
*D.Parameter $\beta$ and $\omega$ evaluation*

The effects of different values of parameter $\beta$ and $\omega$ on the model performance were evaluated on different datasets.

In the first stage, the performance of the model improved continuously as the parameter increased from 0 to 0.02. The performance of the model began to decline at 0.03 and then increased from 0.03 to 0.04. This result indicates that the original face image contains important information. The effect of parameter $\beta$ on the model performance for different data sets is shown in Figure 4(a).




(a)




(b)

Fig. 4: (a) The performance evaluation of parameter $\beta$ on different datasets; (b) The performance evaluation of the distance $\omega$ on different datasets.

In the second stage, we used the optimal settings from the first stage to evaluate the effects of different distances $\omega$ (1.0, 2.0, 3.0, and 4.0) on the model performance. The optimal performance was observed for $\omega$=3.0 (Fig. 4(b)). This result shows that a larger distance results in better differentiation between classes. However, if the distance is too large, over-fitting of the model may occur, resulting in training and performance degradation.

*E. Comparison of the RAMN with other methods*

The RAMN was compared with other methods on the FER2013, Oulu-CASIA, CK+, and RAF-DB datasets.

The confusion matrices in Fig. 5 show that the RAMN with the loss function has better performance than the baseline RAMN on each dataset. The most significant improvement is observed for the uncontrolled dataset RAF-DB. In addition to a higher overall recognition rate of the RAMN with the loss function, the accuracy of the "happy" and "surprised" classes has been significantly improved, indicating that these

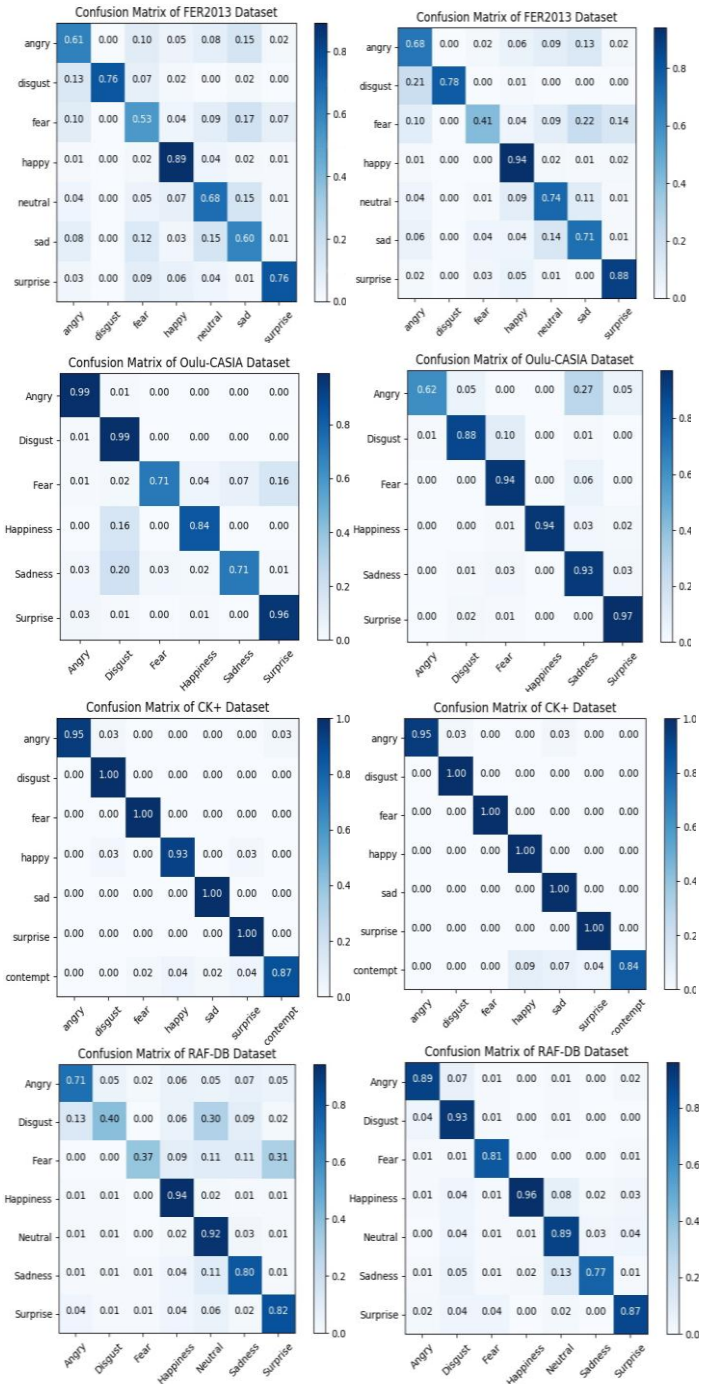



Fig. 5. The confusion matrices of the baseline RAMN (left) and RAMN with a loss function (right) on the FER2013, Oulu-CASIA, CK+, and RAF-DB datasets.

expressions have strong features and are easily learned. The combined use of the R-Loss and C-Loss functions ensures excellent learning performance for the facial texture features and a good balance between the intra-class and interclass differences in facial expressions.

In addition, we conducted experiments on a cross-dataset. Fig. 6 shows examples of the results of the baseline RAMN model and the RAMN with the loss function. We used multiple human face images selected from the AffectNet dataset, including frontal, multi-pose, and occlusion conditions, and the model was trained on the RAF-DB dataset. Under these complex conditions, the recognition results of the baseline RAMN model are not satisfactory. The performance, indicating that the model has good practical application ability and is reliable and practical for analyzing


Fig. 6: Examples of the results of the baseline RAMN model and the RAMN with the loss function on a cross-dataset.

real scenes.

Table I lists the recognition effects of the RAMN and other methods on the FER2013 dataset. The RAMN method has achieved good experimental results. The reason is the combination of the two loss functions for optimizing the model. During the training process, the local facial features are considered, and the differences within and between classes are effectively balanced. The model developed by Zhang *et al*. achieved a better result of 75.10% on FER2013, but they used an additional dataset, whereas we only used the FER2013 dataset.

TABLE I
COMPARISON OF THE ACCURACY OF RAMN AND OTHER METHODS ON THE FER2013 DATASET

| Methods | Acc(%) |
|---|---|
| Mollahosseini*et al* | 66.40 |
| Kim *et al* | 70.58 |
| Tang | 71.16 |
| Connie *et al*[33] | 73.70 |
| Zhang *et al*[34] | 75.10 |
| RAMN(BASELINE) | 70.15 |
| RAMN($\beta$) | 74.49 |
| RAMN( $\beta$ AND $\omega$) | **75.14** |

Table II shows accuracies of the RAMN and other methods on the CK+ and Oulu-CASIA datasets. All deep learning methods have high accuracy. For example, DeRL uses the conditional GAN (CGAN) to generate neutral expressions of the input face. The middle layer of the generator can determine the face expressions, and this information is used to obtain identity-invariant representations of the face expression. The RAMN method has the highest accuracy on the CK+ data set and the second-highest accuracy on the Oulu-CASIA dataset, indicating the excellent execution of the proposed C-loss. In addition, all methods achieved good effects on the CK+ dataset because the images are frontal face views acquired under laboratory conditions, allowing for good discrimination of the expressions.

Table III compares the accuracy of RAMN with that of the other methods on the RAF-DB dataset. RAF-DB is the latest facial expression database, with images of basic expressions and composite expressions. This table demonstrates the consequences of experiments using the basic expressions.

TABLE II
COMPARISON OF THE ACCURACY OF RAMN AND OTHER METHODS ON THE CK+ AND OULU-CASIA DATASETS

| Methods | Acc(%) | |
|---|---|---|
| | CK+ | Oulu-CASIA |
| LBP-TOP | 88.99 | 68.13 |
| AUDN | 93.70 | - |
| IL-CNN | 94.35 | 77.29 |
| DeRL[35] | 97.30 | **88.00** |
| RAMN(BASELINE) | 96.08 | 86.06 |
| RAMN($\beta$) | 96.73 | 86.92 |
| RAMN($\beta$ AND $\omega$) | **97.85** | 87.79 |

The DLP-CNN uses a partial retention loss to train the network, the gACNN uses a portion of the face area and the entire image to train the network, and RAN uses the self-attention module and the relational attention module to train the network. The accuracy rate of the proposed RAMN on the RAF-DB dataset is 86.87%, which is 2.74% and 1.80% higher than those of DLP-CNN and gACNN, respectively, and only 0.03% less than that of RAN.

TABLE III
COMPARISON OF THE ACCURACY OF RAMN AND OTHER METHODS ON THE RAF-DB DATASET

| Methods | Acc(%) |
|---|---|
| DLP-CNN[36] | 84.13 |
| gACNN[37] | 85.07 |
| RAN | **86.90** |
| RAMN(BASELINE) | 82.83 |
| RAMN($\beta$) | 85.04 |
| RAMN($\beta$ AND $\omega$) | **86.87** |

The accuracy of RAMN was 3.38% higher on the CK+ dataset and 1.73% higher on the Oulu-CASIA dataset than that of the baseline RAMN. For the FER2013 and RAF-DB datasets, the accuracy of RAMN was 4.99% and 4.04% higher, respectively, than that of the baseline RAMN, indicating that the proposed method is more suitable than the baseline method for processing data in the natural state. The accuracy was significantly higher after adding the R-Loss function, demonstrating that the network learned useful features in the first stage. After adding the C-Loss, the overall accuracy is improved by about 1.12%, showing the effectiveness of the loss function for balancing the within and

between class differences.

## V. Conclusion

A FER model based on the RAMN was proposed. Two loss functions were incorporated, and the facial images were cropped to conduct a fine-grain analysis. The results showed that cropping into 5 sub-images was optimal. The effect of the two loss functions was evaluated using four datasets. The experiments showed that the proposed RAMN has superior recognition accuracy and generalization ability and outperformed most facial expression models. Our next work will be to improve the C-Loss function because it did not enhance the performance of the network as well as the R-Loss function.

## References


[1] Giambelluca, F. Luis et al. "Scorpion detection and classification systems based on computer vision and deep learning for health security purposes." *ArXiv Preprint ArXiv*:2105.15041, 2021.

[2] L. Pablo, and M. V. Gerven. "Neuroscience-Inspired Perception-Action in Robotics: Applying Active Inference for State Estimation, Control and Self-Perception." *ArXiv Preprint ArXiv*:2105.04261, 2021.

[3] Y. Edouard et al. "RED : Looking for Redundancies for Data-Free Structured Compression of Deep Neural Networks." *ArXiv Preprint ArXiv*: 2105.14797, 2021.

[4] G. Gianluca and M. Piastra. "Conditional Deep Convolutional Neural Networks for Improving the Automated Screening of Histopathological Images." *ArXiv Preprint ArXiv*:2105.14338, 2021.

[5] C. Rathgeb, et al. "Deep Face Fuzzy Vault: Implementation and Performance," *ArXiv Preprint ArXiv: 2102.02458*, 2021.

[6] H. Walid, et al. "Deep and Shallow Covariance Feature Quantization for 3D Facial Expression Recognition." *ArXiv Preprint ArXiv*:2105.05708, 2021.

[7] G. Baris et al. "Fast-GANFIT: Generative Adversarial Network for High Fidelity 3D Face Reconstruction." *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[8] Chen, Junhua, et al. "Lung Cancer Diagnosis Using Deep Attention Based on Multiple Instance Learning and Radiomics." *ArXiv: Image and Video Processing*, 2021.

[9] D. C. Yan, et al. "MUSE: Multi-Faceted Attention for Signed Network Embedding." *ArXiv: Social and Information Networks*, 2021.

[10] Gruel, Am'elie and J. Martinet. "Bio-inspired visual attention for silicon retinas based on spiking neural networks applied to pattern classification." *ArXiv Preprint ArXiv*:2105.14753 , 2021.

[11] S. Ayan et al. "HIT: A Hierarchically Fused Deep Attention Network for Robust Code-mixed Language Representation." *ArXiv Preprint ArXiv*:2105.14600 , 2021.

[12] J. Yang and H. S. Yu. "Attention-oriented Brain Storm Optimization for Multimodal Optimization Problems." *ArXiv Preprint ArXiv*:2105.13095 , 2021.

[13] H. Y. Gong, et al. "Abusive Language Detection in Heterogeneous Contexts: Dataset Collection and the Role of Supervised Attention." *ArXiv Preprint ArXiv*:2105.11119 , 2021.

[14] J. H. Xia, et al. "An Efficient Multitask Neural Network for Face Alignment, Head Pose Estimation and Face Tracking." *ArXiv: Computer Vision and Pattern Recognition*, 2021.

[15] Tommola, Janne, et al. "Real Time System for Facial Analysis." *ArXiv Preprint ArXiv*: 1809.05474, 2018.

[16] Karmakar, Arnab, and D. Mishra. "Pose Invariant Person Re-Identification Using Robust Pose-Transformation GAN." *ArXiv Preprint ArXiv*:2105.00930, 2021.

[17] Hass, D. Ryan, et al. "The Subgrid Scale Pressure Field of Scale-Enr iched Large Eddy Simulations Using Gabor Modes." *ArXiv Preprint ArXiv:2012.15021*, 2020.

[18] Z. R. Huang. "CN-LBP: Complex Networks-Based Local Binary Patterns for Texture Classification." *ArXiv: Image and Video Processing*, 2021.

[19] B. Fasel, "Robust Face Analysis Using Convolutional Neural Networks," *Object Recognition Supported by User Interaction for Service Robots*, vol. 2, 2012, pp. 40-43.

[20] M. Y. Liu, et al. "Au-inspired deep networks for facial expression feature learning," *Neurocomputing* 159 (2015): pp. 126-136.

[21] Ardeshir, Navid et al. "Support vector machines and linear regression coincide with very high-dimensional features." *ArXiv Preprint ArXiv*:2105.14084, 2021.

[22] F. V. Massoli, et al. "MAFER: A Multi-Resolution Approach to Facial Expression Recognition." *ArXiv: Computer Vision and Pattern Recognition*, 2021.

[23] J. B. Xia, et al. "An Efficient Multitask Neural Network for Face Alignment, Head Pose Estimation and Face Tracking." *ArXiv: Computer Vision and Pattern Recognition*, 2021.

[24] Gromada, Xavier, et al. "Le Régulateur Du Cycle Cellulaire E2F1 Module l'expression Du Récepteur Au GLP-1 Dans La Cellule Beta Pancréatique," *Diabetes & Metabolism*, vol. 43, no. 2, 2017.

[25] V. A. Sablina, and A. D. Sergeeva. "METHODS OF THE FACIAL MICRO-EXPRESSION RECOGNITION: AN OVERVIEW." *Modern Technologies in Science and Education MTSC-2019*, 2019.

[26] J. Wan, et al. "Robust Facial Landmark Detection by Cross-Order Cross-Semantic Deep Network." *Neural Networks*, vol. 136, 2021, pp. 233–243.

[27] Menard, G., et al. "Le Modèle Larvaire Galleria Mellonella : Un Modèle Alternatif Pour Étudier l'expression Des ARNs Bactériens Au Cours de l'infection à Staphylococcus Aureus," *Medecine Et Maladies Infectieuses*, vol. 48, no. 4, 2018.

[28] C. Szegedy, W. Liu, et al. "Going Deeper with Convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1-9.

[29] S. Karen, and Z. Andrew. "Very deep convolutional networks for large-scale image recognition," *Computer Vision and Pattern Recognition*, 2014.

[30] M. Y. Liu, et al. "Learning Expressionlets on Spatio-Temporal Manifold for Dynamic Facial Expression Recognition," '*CVPR*' *14 Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1749-1756.

[31] K. Wang, et al. "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing* 29 (2020): pp. 4057-4069.

[32] Y. C. Tang. "Deep learning using linear support vector machines," *arXiv preprint arXiv:1306.0239* (2013).

[33] T. Connie, et al. "Facial Expression Recognition Using a Hybrid CNN–SIFT Aggregator," *International Workshop on Multi-Disciplinary Trends in Artificial Intelligence*, 2017, pp. 139-149.

[34] Z. P. Zhang, et al. "Learning Social Relation Traits from Face Images," *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3631-3639.

[35] H. Y. Yang, C. Umur, and L. J. Yin, "Facial expression recognition by de-expression residue learning," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

[36] S. Li and W. Deng. "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, 28(1):356-370, Jan 2019.

[37] Y. Li, et al. "Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, 2019, pp. 2439-2450.