Machine Intelligence Research
www.mi-research.net

21(2), April 2024, 331-348
DOI: 10.1007/s11633-023-1417-9

Adaptively Enhancing Facial Expression Crucial Regions via a Local Non-local Joint Network

Guanghui Shi[1] Shasha Mao[1] Shuiping Gou[1] Dandan Yan[1]
Licheng Jiao[1] Lin Xiong[2]

[1] School of Artificial Intelligence, Xidian University, Xi'an 710000, China

[2] Applied Research Laboratory, SenseTime, Xi'an 710000, China

Abstract: Facial expression recognition (FER) is still challenging due to the small interclass discrepancy in facial expression data. In view of the significance of facial crucial regions for FER, many existing studies utilize the prior information from some annotated crucial points to improve the performance of FER. However, it is complicated and time-consuming to manually annotate facial crucial points, especially for vast wild expression images. Based on this, a local non-local joint network is proposed to adaptively enhance the facial crucial regions in feature learning of FER in this paper. In the proposed method, two parts are constructed based on facial local and non-local information, where an ensemble of multiple local networks is proposed to extract local features corresponding to multiple facial local regions and a non-local attention network is addressed to explore the significance of each local region. In particular, the attention weights obtained by the non-local network are fed into the local part to achieve interactive feedback between the facial global and local information. Interestingly, the non-local weights corresponding to local regions are gradually updated and higher weights are given to more crucial regions. Moreover, U-Net is employed to extract the integrated features of deep semantic information and low hierarchical detail information of expression images. Finally, experimental results illustrate that the proposed method achieves more competitive performance than several state-of-the-art methods on five benchmark datasets.

Keywords: Facial expression recognition, deep neural network, multiple network ensemble, attention network, facial crucial regions.

Citation: G. Shi, S. Mao, S. Gou, D. Yan, L. Jiao, L. Xiong. Adaptively enhancing facial expression crucial regions via a local non-local joint network. *Machine Intelligence Research*, vol.21, no.2, pp.331–348, 2024. http://doi.org/10.1007/s11633-023-1417-9

1 Introduction

Emotion is a complex state that integrates people's feelings, thoughts and behaviors[1], and facial expression is one of the most direct signals to communicate their innermost thoughts. Therefore, facial expression recognition (FER)[2–6] has attracted the attention of many researchers due to its important role in many practical application fields, such as human-computer interaction, recommendation system, patient monitoring, etc. In general, facial expressions are encoded into facial action units through facial action coding system[7, 8], and any expressions can be described through a set of facial action units. Some facial action units are crucial for FER[9], such as those located in regions around the eyes and mouth, since they have more obvious actions than other facial regions (such as cheek and forehead). In the following parts, we regard these crucial facial action units as facial crucial regions (FCRs). Fig. 1 illustrates facial crucial regions of two facial images (ID1 and ID2) from six expressions, respectively. From Fig. 1, it is found that the FCRs are more discriminative to determine the expression category of a facial image[10].

In view of the significance of FCRs, many studies[11–14] have been proposed based on applying the information of facial local regions, where the facial landmarks are employed as the prior information to obtain facial crucial regions. However, the information of facial landmarks is obtained by manual annotation. Early, most of FER studies[15–17] focused on lab-collected expression datasets, such as CK+[18], MMI[19], JAFFE[20], Oulu-CASIA[21]. For lab-collected datasets, facial expressions images were collected from several or dozens of individuals under similar conditions (such as illumination, angle, posture, etc.), generally with a few uncontrollable factors. Thus, it is easily achieved to manually annotate the landmark of FCRs for lab-collected datasets.

However, compared with the lab-controlled datasets, the wild expression datasets[22] are collected under more complex and uncontrollable conditions, such as RAF-DB[23], AffectNet[24], EmotionNet[25], etc. For the wild expression datasets, especially including a vast of images, it

Research Article
Manuscript received on November 30, 2022; accepted on January 18, 2023; published online on January 11, 2024
Recommended by Guest Editor Lun-Tian Mou
Colored figures are available in the online version at https://link.springer.com/journal/11633

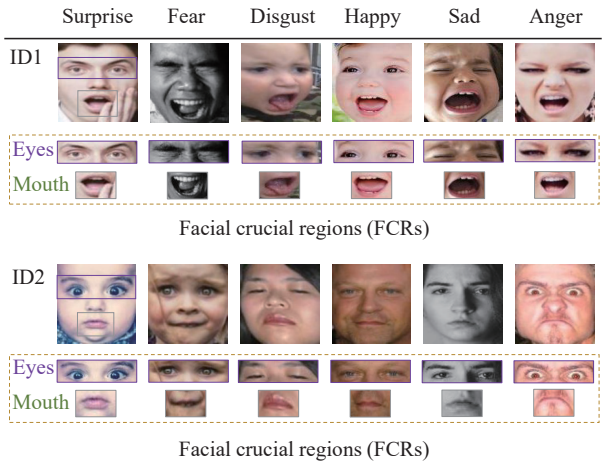


Fig. 1 An illustration of facial crucial regions from six expressions, where two facial images (ID1 and ID2) from RAF-DB[23] are shown for each expression. The regions around eyes and mouths are cropped as examples of FCRs in the purple box and the green box, respectively.

is very complicated and time-consuming for manually annotating FCRs. Moreover, the postures of different faces vary greatly on the wild database. One simple change in facial postures can cause multiple pixel deviations at the image level. Fig. 2 gives an example of the landmarks moving with the change of postures, where two expression images and their landmarks are from RAF-DB dataset[23]. From Fig. 2, it is observed that 68 landmark points of subimage (a) are different from subimage (b), and the landmarks are greatly shifted from (a) to (b), shown as subimage (c). This implies that the position of FCRs varies with the change in facial postures. Inevitably, it increases the complexity of manually annotating landmarks for FER, especially for wild datasets with vast numbers of images. In view of this, it is important to consider whether the significance of FCRs or their features could be spontaneously enhanced in the training of deep FER, without any prior information, such as landmarks of FCRs.

On the other hand, there exists a problem that some FCRs from different expression categories are similar, whereas some FCRs from the same category are very different. From Fig. 1, it is obviously seen that the FCRs (including mouths) of ID1 from six expressions are similar with opening the mouth, which is absolutely different from ID2 with closing the mouth. Similarly, for the crucial regions including eyes, ID1 and ID2 from the category (Fear) are different, whereas ID1 from the category (Surprise) and ID2 from the category (Anger) are similar. This illustrates that the FCRs of expression images belonging to the same category may be very different, but FCRs from different categories are similar. Distinctly, it is insufficient that only local information of facial expressions is utilized to construct one effective model for FER, especially for the wild dataset. Hence, it is still important to utilize the global information of the facial expression while FCRs are enhanced in deep facial expression recognition.

Fig. 2 Schematic diagram of the pixel deviations at the image level when posture changing. To demonstrate this change, we measured the movement of 68 landmark points on faces with different postures and the same identity. In (a) and (b) from RAF-DB[23], 68 landmark points are marked with a green cross, and (c) shows the movement of 68 landmark points.

Based on the above analyses, we propose a new method of facial expression recognition in this paper, which constructs a local non-local joint network to adaptively enhance the facial crucial regions in the process of deep feature learning, shortened for LNLAttenNet. In LNLAttenNet, the local and non-local information of facial expressions are simultaneously considered to construct two parts of the network: a local multi-network ensemble and a non-local attention network, and then the generated local and non-local feature vectors are integrated and jointly optimized in feature learning. Specifically, the attention weights obtained by the non-local part are regarded as the significance of facial local regions and fed into the local multi-network ensemble system to combine multiple local networks. Interestingly, we find that some FCRs can be automatically enhanced in the process of deep feature learning by the proposed method. Moreover, U-Net is employed to generate feature maps where each pixel has a large receptive field and the local region also contains global information. Fig. 3 shows a simple view of LNLAttenNet. From Fig. 3, it is obvious that some crucial regions are given higher weights by LNLAttenNet, such as the 5th patch around the left eye (0.112 3) and the 10th, 11th and 14th patches around the mouth (0.088 7, 0.107 3 and 0.129 8), which illustrates that some crucial regions are effectively enhanced by LNLAttenNet. Note that w_i is the non-local attention weight corresponding to the i-th local region and the initial weights are equal. More detailed descriptions will be introduced in the following sections.

Compared with state-of-the-art methods, our contributions are mainly three points:

1) We propose LNLAttenNet to automatically enhance facial crucial regions in deep feature learning by utilizing the local and non-local information of facial expressions simultaneously. To the best of our knowledge, this is the first study on how to explore and enhance the FCRs in CNNs for FER, where FCRs are automatically enhanced without any prior information for facial crucial regions or landmarks. It effectively improves the problem that it is difficult to annotate the facial landmarks of the

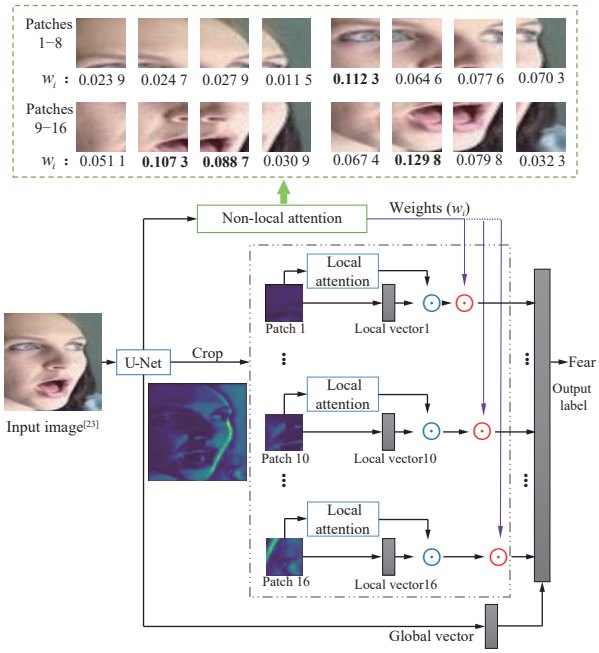


Fig. 3 A simple view of the proposed model (LNLAttenNet). The part in the green dotted box shows the global weights corresponding to 16 local regions (from Patch 1 to Patch 16) obtained by LNLAttenNet, and the part under the green dotted box is a simple framework of LNLAttenNet.

wild facial datasets.

2) In LNLAttenNet, an attention mechanism is introduced to construct a non-local attention network that explores the significance of local regions for FER from a global perspective of facial expression. The obtained attention weights corresponding to local regions are fed into the local multi-network ensemble system to integrate multiple local features, and then the integration of features obtained by multiple local networks is jointly optimized with the facial global feature.

3) Experimental results demonstrate that FCRs can be enhanced in deep feature learning by LNLAttenNet, which validates that FCRs are more discriminative local regions for FER. Moreover, it also implies that the deep FER model can spontaneously focus on some crucial regions in the training process, which probably brings a new inspiration for designing deep FER methods.

The rest of this manuscript is organized as follows. Section 2 first introduces related works about deep facial expression recognition. Second, Section 3 introduces the details of the proposed method. Then, experimental results and analyses are demonstrated to validate the performance of the proposed method in Section 4. Finally, Section 5 provides the conclusion as well as prospects for future work.

2 Related work

Due to the excellent performance of deep learning[26–28], various deep networks have been applied in FER[22], such as VGGNet[29], InceptionNet[30] and ResNet[31]. Based on this, many deep FER methods have been proposed to address different problems. Hu et al.[32] first extended the idea of deep supervision to address FER in the wild. The training of deep CNNs was softer and easier through supervision not only to deep layers but also to intermediate layers and shallow layers, and a fusion structure was constructed where the feature ahead was used for second-level supervision. Acharya et al.[33] thought that second-order statistics (such as covariance) were more suitable to capture the features of twisted facial expressions. In their framework, a mainfold structure was constructed for covariance pooling to obtain a competitive performance for FER. Li and Deng[34] proposed a new deep manifold strategy for multi-label expressions, and their proposed network focused on ambiguous expressions and could learn the discriminative feature that was suitable for cross-database FER.

Considering that facial expression is determined by key regions, Fan et al.[11] utilized the information of facial landmark points to select three subimages around the eyes, mouth and nose. Then, three subimages were encoded by three subnetworks, and the last pooling layer in each sub-network was concatenated with each other, which obtained better recognition performance compared with others. In [35, 36], facial landmark information is used to extract features and generate masks from specific locations to remove the pose variation.

In [37], it was taken into account that there are inevitably labelling errors and deviations between different databases due to the subjectivity of labelling facial expressions. Therefore, when existing methods make use of multiple databases to expand the training set, their performance cannot be continuously improved. To solve this inconsistency between different databases, an IPA2LT framework is proposed to train a model from multiple inconsistent databases and large scale unlabelled images. The IPA2LT essentially constructs the ensemble at the label level. Each image in the model has the same number of labels as the number of data sources, in which only one label is original and others are pseudo. Existing methods for FER have been almost satisfying in analyzing frontal faces but fail to attain good performance on partially occluded faces collected in the wild. Some facial expressions are ambiguous and have multiple labels. Gan et al.[38] proposed a new framework based on CNN with the supervision of soft labels, where hard labels are used to construct soft labels with a novel label-level perturbation. In this framework, soft labels were obtained to eliminate the similarity between faces of different emotions, and multiple basic classifiers were trained and then combined. Moreover, some GAN-based methods have been proposed to generate expressional images for FER[39–41] or usually focus only on generating new facial expression images[42–45]. In [39], a novel approach is proposed to learn facial expressions by extracting the expressive compon-

ent through a de-expression procedure where the corresponding neutral expression is generated by the trained generative model given a facial image with arbitrary expressions. In [42], a user-controllable approach is proposed to generate video clips of various lengths from a single face image and the lengths and types of the expressions are controlled by users.

Li et al.[12] proposed a CNN with an attention mechanism (ACNN) to detect the occlusion of facial regions and paid attention to the most discriminative regions, where the ACNN used the information of 24 facial landmark points to select the key regions at the feature level. Barros et al.[46] investigated emotion-driven attention mechanisms from the view of videos. Wang et al.[47] proposed a two-level attention mechanism to extract emotion-related features, which was based on global information, and did not involve the local regions. Similar to [12, 46, 47], the attention mechanism is also involved in this work, whereas the essence of algorithms is very different. Here, our purpose is to adaptively enhance the significance of facial crucial regions based on the attention weights in feature learning obtained by the non-local attention network from the view of multiple local regions, where the attention weights corresponding to each local region are obtained by the non-local attention network.

3 Local non-nocal joint network for FER

In this paper, we propose a local non-local attention joint network for FER to adaptively enhance more crucial local regions of facial expression, named by LNLAttenNet. The overall framework of LNLAttenNet is visually shown in Fig. 4. In Fig. 4, one facial expression image is used as the initial input instance of the proposed network, and its size is 144×144, as in our implemented experiments.

In LNLAttenNet, U-Net is first employed to extract the feature maps integrating the deep semantic information and the low hierarchical detail information of facial expression images. For the facial expression dataset, when regional integration is carried out[11], the interclass discrepancy is smaller and the intraclass discrepancy is larger, as shown in Fig. 1. The structure of U-Net[48–50], the top-down architecture with lateral connections for introducing details into high-level semantic feature maps, has been proven that local regions in the last few layers have a large receptive field and the global information, which is important and useful for ambiguous object recognition[51, 52]. Therefore, U-Net is beneficial for alleviating the negative impact of the regional integration, but it does not mean that the proposed method is restricted to

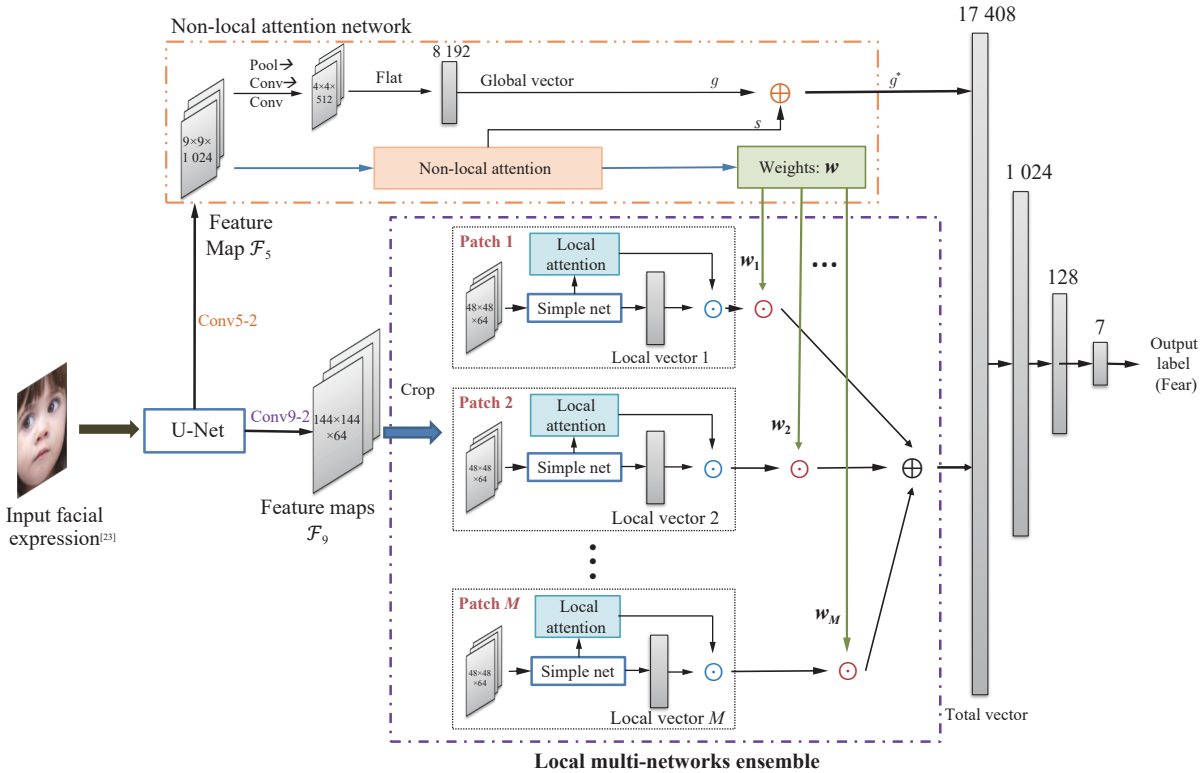


Fig. 4 The framework of the proposed model (LNLAttenNet). LNLAttenNet uses U-Net to generate a feature map with the same resolution as the input image. Then, its feature map (Conv9-2) is cropped into M local patches to construct the local multinetworks ensemble model, where each patch is used to generate an individual network based on the structure of simple net. Feature map (Conv5-2) is used to construct the global attention network. Finally, the global and local features are integrated based on the global weights, and then three fully connected layers follow.

U-Net. One model with a similar structure to U-Net can be employed in our proposed method, such as feature pyramid network (FPN)[51].

As shown in Fig. 4, facial expression images are inputted to the proposed model. By U-Net, two different feature maps are generated for the initial input image, located in the last layer (Conv9-2) and the intermediate layer (Conv5-2) of U-Net. In the following parts, we use $\mathcal{F}_5$ and $\mathcal{F}_9$ to express the feature maps from Conv5-2 and Conv9-2 of U-Net, respectively. Then, the generated feature maps $\mathcal{F}_5$ and $\mathcal{F}_9$ are utilized to construct two parts of LNLAttenNet, where the map $\mathcal{F}_5$ is utilized as the input to construct the non-local part (the non-local attention network) and the map $\mathcal{F}_9$ is employed as the input to construct the local part (the local multi-networks ensemble system). In the local part, an ensemble of multiple networks is applied to generate and integrate multiple individual networks corresponding to different facial local regions. By the non-local attention network, an attention weight w_i $(i = 1, \cdots, M)$ is obtained corresponding to the i-th local region of the facial expression, and then the vector $\boldsymbol{w}$ $([w_1, \cdots, w_M]^{\mathrm{T}})$ is used as the weights of multiple local networks to combine M local vectors and boost the significance of local regions in the process of deep feature learning. Finally, the non-local attention network and the local ensemble network are jointly optimized by integrating local and non-local features in three fully connected layers of LNLAttenNet. More detailed descriptions of the proposed method will be introduced as follows.

3.1 Non-local attention network

For facial expression recognition, there are the small interclass discrepancy and the large intraclass discrepancy on expression images, as shown in Fig. 1. Therefore, facial crucial regions are regarded as more discriminative regions that determine the categories of facial expression, such as regions around the mouth (eyes) rather than the cheek. However, it is difficult to estimate which regions are more crucial without assistance from manually annotated crucial points. Based on this, we construct the non-local attention network to automatically mine more discriminative regions from the whole facial expression, visually shown in the box with orange dotted lines of Fig. 4.

In Fig. 4, the feature map $\mathcal{F}_5$ (Conv5-2) is generated by U-Net as the global information of the facial image to construct the non-local attention network. Conv5-2 has a minimum resolution and the maximum receptive field, which means that $\mathcal{F}_5$ is not affected by each local patch but implicitly contains the relationship between local patches. It is useful to mine more crucial regions based on the global information from the whole face.

Inspired by [53, 54], we construct a non-local attention model based on three branches, as shown in Fig. 5. First, the input is map $\mathcal{F}_5$, which contains the global information of facial expression in Fig. 5. Based on $\mathcal{F}_5$, three feature maps $\mathcal{Q}$, $\mathcal{K}$ and $\mathcal{V}$ are generated by one convolution layer and one pooling layer. Note that the three maps have a special resolution[1] of $n \times n$ in this model, where $M = n^2$ and M is the number of cropped local regions. Then, the maps $\mathcal{Q}$ and $\mathcal{K}$ are reshaped as $\boldsymbol{Q}^*$ and $\boldsymbol{K}^*$ respectively, as shown in Fig. 5, and a multiplication operation is followed to obtain a matrix $\boldsymbol{R}$ that reflects the correlation among local regions. Compared with [53, 54], the relevance of each region (patch) in LNLAttenNet is not as strong as each frame in the video or each word in the sentence, and thus, L_1 normalization is adopted to limit the sum of each row of R to 1 instead of the softmax function. Finally, a vector is calculated by averaging each column of the correlation matrix $\boldsymbol{R}$, regarded as the non-local attention weights $\boldsymbol{w}^g$ assigned to M local regions.

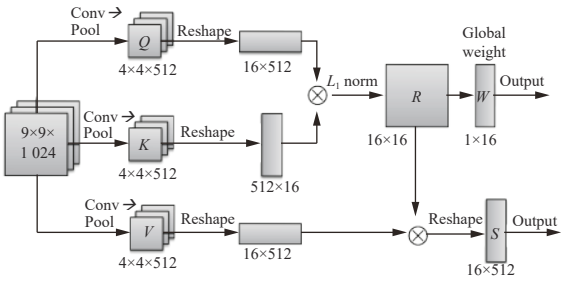


Fig. 5 Overview of the non-local attention model

Furthermore, the map $\boldsymbol{V}$ is reshaped as $\boldsymbol{V}^*$, and the feature vector $\boldsymbol{s}$ is obtained by multiplying $\boldsymbol{V}^*$ by the correlation matrix $\boldsymbol{R}$, which is the self-attention form in [53, 54]. To make the matrix $\boldsymbol{R}$ reflect the correlation among local regions, $\boldsymbol{s}$ is flattened and added to the non-local vector $\boldsymbol{g}$ (shown in Fig. 4). Meanwhile, a function is given to trade off two vectors $\boldsymbol{g}$ and $\boldsymbol{s}$, shown as

$$\boldsymbol{g}^* = (1 - \alpha) \times \boldsymbol{g} + \alpha \times flat(\boldsymbol{s}) \tag{1}$$

where $\boldsymbol{g}^*$ expresses the new non-local vector and α is the hyperparameter to adjust the ratio of $\boldsymbol{s}$. In experiments, we will give an analysis for the parameter α.

3.2 Local multi-networks ensemble

The feature map ($\mathcal{F}_9$) is employed as the input to construct the local multi-networks ensemble, as shown in Fig. 4. The reason for using the map $\mathcal{F}_9$ is that each pixel is of the large receptive field and the rich semantic information in Conv9-2, where $\mathcal{F}_9$ has the same resolution as the initial input image. In the part of local multi-networks ensemble, the feature map $\mathcal{F}_9$ is first divided into M patches (including different local regions) with the

[1] This special resolution is set to expediently calculate the correlation between each patch. For example, when the number of cropped local regions is set as 16 ($M = 16$) in our experiments, the special resolution is 4×4 ($n = 4$), as shown in Fig. 5.

same dimension (set as 48 × 48 × 64 in our experiments). Then, M patches are trained by the sample network to generate M individual networks $\{\mathcal{IN}_1, \cdots, \mathcal{IN}_M\}$, respectively. The basic structure of the simple network is shown in Fig. 6, composed of six convolution layers and three pooling layers. Specifically, for each individual network, the local attention mechanism is added to enhance the feature vector of each local region. Finally, M local feature vectors are combined with the non-local attention weights obtained by the non-local attention network.

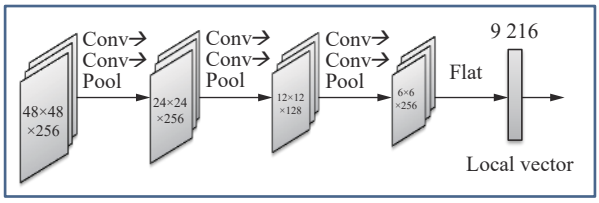


Fig. 6 Structure of the simple network

Local attention. In practice, the useful information is decreased when partial regions in one patch are missed or obscured. This means that less attention should be given to them. In view of this, a local attention mechanism is adopted in each individual network to weaken the significance of useless regions. The local attention model is encoded by four convolution layers and two fully connected layers, and its structure is shown in Fig. 7. Note that two convolution layers are not padded to reduce the computational complexity. In the local attention model, its input is the output of the last pooling layer in SimpleNet, and its output is one value between 0 and 1 obtained via the sigmoid function, regarded as the local attention weight w_i^l of each individual network, which represents the amount of information in each local patch that can flow to the next level. If the facial local region is obscured or missed, the information that it contains for expression recognition will be reduced, and then the weight value of the local attention is also reduced to alleviate the effect of patches including the obscured region. Furthermore, the weights are multiplied by the corresponding local vector as the output feature of each local network. More visual illustrations can be found in the experiments section.

Combination of multiple local networks. According to the non-local attention weights $\boldsymbol{w}^g$ and the local attention weights $\boldsymbol{w}^l$, the local feature vectors given by M individual networks $\{\mathcal{IN}_1, \cdots, \mathcal{IN}_M\}$ are aggregated by the formula

$$\boldsymbol{f}_{en} = \sum_{i=1}^{M} w_i^g \times w_i^l \boldsymbol{f}_i \tag{2}$$

where $\boldsymbol{f}_{en}$ expresses the ensemble feature vector, $\boldsymbol{f}_i$ expresses the feature vector given by $\mathcal{IN}_i$ corresponding to the i-th local region, w_i^g is the non-local attention weight of the i-th local region, and w_i^l expresses the local attention weight of the i-th local region. In the experiments, we analyse the number M of local patches.

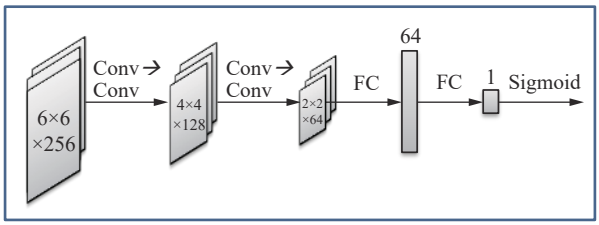


Fig. 7 Overview of the local attention

3.3 Joint optimization of LNLAttenNet

In Fig. 4, the non-local feature vector $\boldsymbol{g}^*$ is produced by the non-local attention network, and the local vector $\boldsymbol{f}_{en}$ is obtained by the local multinetwork ensemble. Inspired by [55], we think that the global information of an input image is essential, and each local patch can obtain a large receptive field and global information by embedding U-Net, which makes it easier to classify similar patches of facial expressions of different categories. Moreover, Conv5-2 is encoded to a global vector with 8 192 dimension by two convolution layers and one pooling layer. Then, the non-local vector $\boldsymbol{g}^*$ is concatenated with the local vector $\boldsymbol{f}_{en}$ to obtain the total vector as the feature of the first fully connected layer and is jointly optimized, and the dimension of the integrated feature vector is 17 408, as shown in Fig. 4. In LNLAttenNet, three fully connected layers are implemented, and the loss function is formulated as

$$L = loss_{entropy} + \gamma loss_{l2} \tag{3}$$

where $loss_{entropy}$ expresses the cross entropy loss, $loss_{l2}$ is the $l2$ regularization loss, and γ is the hyperparameter controlling the balance between two losses. The cross entropy is calculated as follows:

$$loss_{entropy} = \frac{1}{N} \sum_{n=0}^{N-1} \sum_{c=0}^{C-1} I(l_n = c) \times \log(p_n^i) \tag{4}$$

where C is the number of categories, N is the number of input images, and I is the function that determines whether the input is correct. p_n^i is the i-th component of the output of the last softmax layer of the n-th image, and l_n is the label of the n-th input image. The $l2$ regularization loss is computed by $loss_{l2} = \lambda \times ||W||^2$, where W is the parameters of our model and λ is set as 0.000 1 in the following experiments.

4 Experiments and analyses

In this section, we will validate the performance of the proposed method from several items: 1) the performance

comparison with state-of-the-art methods on benchmark datasets, 2) the analysis of non-local attention, 3) the visualization of local attention, 4) the change of the parameter α, 5) the performance of LNLAttenNet with different M, and 6) the analyses for overlapped pixels between local regions.

4.1 Databases and setups

In the experiments, we employ five FER datasets to evaluate the performance of LNLAttenNet: RAF-DB[23], SFEW[56], AffectNet[24], CK+[18] and MMI[19].

- **RAF-DB** contains 29 672 facial images downloaded from the Internet. For the RAF-DB dataset, the facial landmarks are manually annotated via the crowdsourcing method with basic or compound expressions. In the experiments, we use the basic database including 12 271 training and 3 068 testing images.
- **SFEW** contains the statistical images selected from the movie clips with spontaneous expressions, where the labels of the training set and validation set are given. Therefore, 958 training images are used as the training set and a total of 436 validation images are used as the testing set in experiments.
- **AffectNet** contains 450 000 images with 10 categories, where each image is annotated by one volunteer. In the experiments, we use 287 401 images with neutral and six basic emotions, where 283 901 images are selected as the training set and 3 500 images are selected from the validation set as the testing set.
- **CK+** contains 593 sequences from 123 volunteers, where 309 sequences have been annotated with six basic emotions. The emotion in each sequence goes from neutral to peak and then to neutral again. In view of this, we select the first frame of each sequence with the label of neutral and the peak frame of each sequence with the target label to generate 618 experimental images.
- **MMI** is recorded from 30 objects with rich details of annotations, and 398 images are generated by selecting the first frame of each sequence with the label of neutral and one peak frame of each sequence.

For the RAF-DB and SFEW datasets, their training sets are directly used to train the model, and testing sets are used to evaluate the performance. For the AffectNet dataset, its training set is used to train the model, and its validation set is used as the testing set, since the testing set of AffectNet is not given the annotated labels[24]. For the CK+ and MMI datasets, we adopt the fivefold cross-validation scheme to evaluate the recognition performance to make a fair comparison with other methods.

Additionally, to fairly compare with the state-of-the-art methods of FER, we initialize the parameters of U-Net by the Xavier initializer rather than pretraining. In the experiments, the original images are resized to 144×144, and the training images are augmented by standard approaches, such as image flips and random cropping. The number M of local regions is set as 16, and each patch (local region) overlaps approximately 16 pixels with its adjacent patches, and the parameter α is set as 0.7 in (1). The size of the epoch is set to 24, the initial learning rate is 0.000 3, and the weight decay is set as 0.95 for each epoch. All experiments are implemented on the framework of TensorFlow and GTX 2080Ti with 11 G memory.

4.2 Comparisons with state-of-the-art methods

To validate the performance of the proposed method, we first compare it with eight state-of-the-art methods on five datasets. Eight compared methods are VGG16[29], DLP-CNN[23], NAL[57], Soft-CNN[38], CenterLoss[58], gACNN[12], LDL-ALSG[59] and IPA2LT[37], where VGG16 is applied as the baseline method in the experiments.

- **DLP-CNN**[23] decomposes the image structurally rather than spatially into regions (parts) that are discriminative for matching. According to the representations over the regions, it aggregates discriminative features for classification.
- **NAL**[57] utilizes a noise adaptation layer to address the problem of noise labels.
- **Soft-CNN**[38] fuses the latent label probability distribution predicted by the trained model to obtain soft labels with a novel label-level perturbation strategy.
- **CenterLoss**[58] minimizes the center loss calculated by the distance between each data point and its corresponding class center to reduce the intraclass discrepancy.
- **gACNN**[12] uses 24 facial landmarks as the attention mechanism to conduct multiregion ensemble at the feature level.
- **LDL-ALSG**[59] considers the subjectivity of human annotators and ambiguous expression labels and then leverages the topological information of the labels from related but more distinct tasks, such as AU recognition and facial landmark detection, to explore the label distribution of facial expressions.
- **Ad-Core**[60] proposes an adaptive correlation (Ad-Corre) loss to guide the network towards generating embedded feature vectors with high correlation for within-class samples and less correlation for between-class samples.
- **IPA2LT**[37] employs an inconsistent pseudo annotations framework to solve the inconsistent annotations between different facial expression databases.

Noticeably, IPA2LT[37] applies both RAF and AffectNet as the training set, differently from our method (LNLAttenNet) and other compared methods where only the training set of one dataset is employed to train a model. In LNLAttenNet, both non-local attention and local attention mechanisms are utilized. Thus, we also make a comparison with three special cases of our model: the model without both local and non-local attention

(Model-S), the model with only local attention (Model-Local), and the model with only non-local attention (Model-NonLocal). Table 1 shows the experimental results of 12 models, where the highest accuracy is bold for each dataset. All results are the average of the last 10 epochs.

From Table 1, it is obviously seen that the performance of the proposed method (LNLAttenNet) is superior to all compared methods except LDL-ALSG and IPA2LT on AffectNet, RAF-DB, CK+, MMI and SFEW. In contrast to LNLAttenNet, IPA2LT[37] utilizes two large datasets (RAF and AffectNet) as the training set, which results in its obtaining better performance. However, LNLAttenNet still achieves competitive performance on two datasets (RAF-DB and SFEW) and outperforms IPA2LT on three datasets (AffectNet, CK+ and MMI). Compared with LDL-ALSG[59], LNLAttenNet outperforms on RAF-DB, SFEW and CK+, ties on AffectNet and loss on MMI. In the last column of Table 1, we also show the average accuracies for five datasets given by each method in the last column. LNLAttenNet obtains the highest average accuracy: 74.03%, which illustrates that LNLAttenNet can obtain a more competitive performance of FER on all five datasets than the eight compared methods.

Furthermore, Model-S is inferior to Model-Local, Model-NonLocal and LNLAttenNet, which demonstrates that the attention mechanism is meaningful for improving the performance of FER in our model. Meanwhile, Model-NonLocal is slightly better than Model-Local but obviously inferior to LNLAttenNet, which also demonstrates our model jointly utilizing local and non-local information of facial expression is more effective. In short, the experimental results illustrate that adaptively enhancing the facial crucial regions in feature learning by LNLAttenNet is effective for improving the performance of FER.

Considering that the RAF and AffectNet datasets have a large number of images, we also show the confusion matrices for them in Figs. 8 and 9, respectively. According to the confusion matrices, it is observed that the categories (fear and surprise) are easily distinguishable for RAF-DB (shown in Fig. 8) and the categories (disgust and anger) are easily distinguishable for AffectNet (shown in Fig. 9).

4.3 Analyses of non-local attention

LNLAttenNet adaptively enhances the feature learning of facial crucial regions by jointly optimizing for local and non-local parts, where the non-local attention network is constructed to obtain the global weights $\boldsymbol{w}^g$ of multiple local regions. One purpose of our work is to explore how to automatically enhance the significance of local crucial regions in deep FER, while any landmarks are not given as the prior information of facial crucial regions. Thus, to validate it, we analyse for the weights of 16 local regions obtained by our non-local attention for the RAF-DB dataset.

First, the visualization results from 16 persons are shown in Fig. 10. In Fig. 10, the first and third rows show the original facial expression images, and the second and fourth rows exhibit the matrix (4×4) of the final global weights $\boldsymbol{w}^g$ (16×1) corresponding to 16 local regions. For each matrix, the darker the color is, the higher the weight is. From Fig. 10, it is obvious that some crucial regions obtain higher weights and noncrucial regions obtain smaller weights for each facial expression. For example, the areas including or around eyes are given higher weights for the first person in the first row, where the maximum is given the local region located at the coordinate (2, 2) including eyes. For the sixth person in the first row, four local regions (located at (3, 2), (3, 3), (4, 2),

Table 1 Accuracy (%) of the proposed method (LNLAttenNet) compared with state-of-the-art methods

Methods	AffectNet	RAF-DB	SFEW	CK+	MMI	Average
VGG16[29]	51.11	80.96	54.45	90.37	63.21	68.02
DLP-CNN[23]	54.47	80.89	–	–	–	–
NAL[57]	55.97	84.22	**58.13**	91.20	64.71	70.85
Soft-CNN[38]	56.77	85.20	55.73	–	–	–
CenterLoss[58]	57.37	84.42	56.19	95.48	–	–
gACNN[12]	58.78	85.07	–	97.03	–	–
LDL-ALSG[59]	59.35	85.53	56.50	93.08	**70.49**	72.99
Ad-Core[60]	**59.67**	86.10	57.64	97.13	67.95	73.70
IPA2LT[37]	*55.11*	*86.77*	*58.29*	*91.67*	*65.61*	*71.49*
Model-S	56.26	83.80	54.82	94.14	63.52	70.51
Model-Local	57.63	84.55	56.42	96.44	65.42	72.09
Model-NonLocal	58.09	85.04	55.73	96.63	66.56	72.41
LNLAttenNet	59.28	**86.15**	57.80	**98.18**	68.75	**74.03**

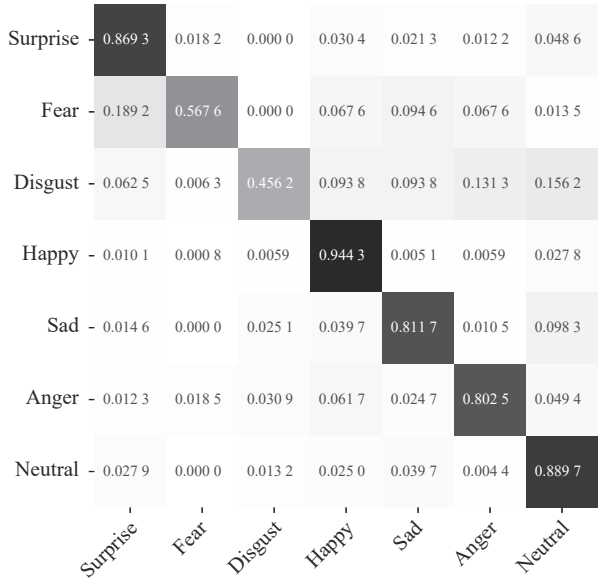


Fig. 8 Confusion matrix of the proposed model (LNLAttenNet) on the RAF-DB database

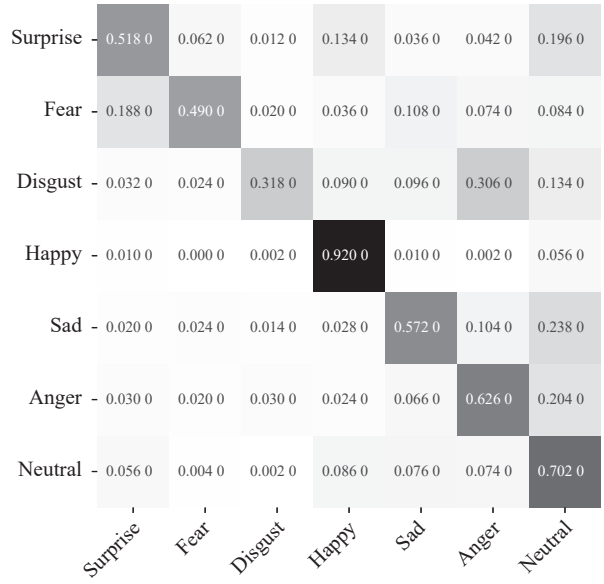


Fig. 9 Confusion matrix of the proposed model (LNLAttenNet) on the AffectNet database

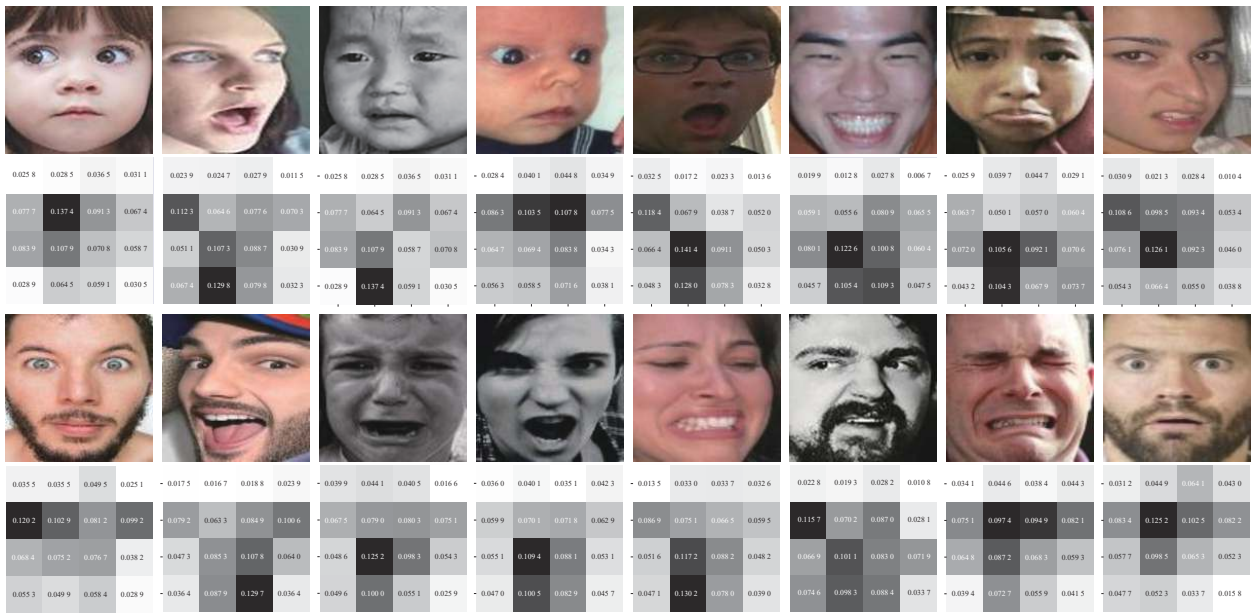


Fig. 10 Non-local weights of 16 local regions of one face in RAF-DB[23] obtained by the proposed model. The first and the third lines show the facial images, and the second and the fourth lines show the non-local weights of 16 local regions corresponding to images.

and (4, 3)), including his mouth, are boosted and given higher weights. In the third and fourth rows, the local regions located around the eyes and mouth are boosted for the second person, and all regions, including the eyes, are given higher weights for the last person. Visually, these enhanced local regions are more discriminative and significant for FER.

From Fig. 10, it is also observed that the location of crucial regions is different for different facial images. However, our network still automatically tracks down more discriminative regions for each different face, without the supervision of any annotated crucial points. Based on this, we conduct an experiment to pursue the change of weights corresponding to each local region in the process of training our model. More visual exhibitions are shown in the Appendix.

4.4 Visualization of local attentions

In the proposed method, local attention is designed to address the problem that local regions are missed or obscured. In this part, the visualization of local attention

will be shown to validate the robustness of the proposed method for faces with missing regions, experimented on the RAF-DB database. Note that the sigmoid function is employed to select the information flowing into the next layer in our local attention model. Fig. 11 shows the visual results of local attention obtained by our method.

In Fig. 11, the 1st and 3rd rows show one original facial image and six obscured images (from the 2nd to 7th columns), and the 2nd and 4th rows show the weights of 16 patches of each facial image obtained by our method. Compared with the result of the original images (shown in the first column of Fig. 11), it is found that the weight is weakened while one patch is obscured and the weights of other patches are unchanged. Note that the weights of some adjacent patches are also decreased with the central patch due to overlapping pixels between two adjacent patches. Practically, the local vector encoded based on one obscured patch is given a small weight, which effectively diminishes the influence of that obscured patch for facial expression recognition. In short, the experimental results illustrate that the proposed method equipped with local attention is more robust for complex facial expression databases in practice.

4.5 Analyses for the parameter α

In the non-local attention network, we formulate (1) to obtain the non-local feature vector $\boldsymbol{g}^*$ based on the global information of facial expression, where the parameter α is used to traffic off the feature vectors $\boldsymbol{g}$ and $\boldsymbol{s}$. In the previous experiments, we set $\alpha = 0.7$. Therefore, we analyse to observe the performance of the proposed method with different values of α in this part. In this experiment, the experimental setups are the same as the above experiments except α, and α is set as {0, 0.1, 0.2, ⋯, 0.9, 1}. Table 2 shows the accuracy under different α for five datasets. From Table 2, it is seen that the accuracy first increases and then decreases with a change in trend while increasing the value of α. According to (1), we obtain $\boldsymbol{g}^* = \boldsymbol{g}$ if $\alpha = 0$ and $\boldsymbol{g}^* = \boldsymbol{s}$ if $\alpha = 1$. Combining the network optimization, it is known that the back propagation in LNLAttenNet has no constraint on $\boldsymbol{s}$ when $\alpha = 0$, which implies that the same effect (or feedback) is given the non-local attention and that each component of the non-local weights $\boldsymbol{w}^g$ should be random in theory. In contrast, $\alpha = 1$ means that the back propagation has no constraint on the global vector $\boldsymbol{g}$, which means that the back propagation in LNLAttenNet has no global information and may result in an extreme result. Actually, as shown in Fig. 12, we also find that the obtained weights ($\boldsymbol{w}^g$) tend to be random under a small α and equal under a large α, which effectively verifies that the effect of α is the same as the above analysis.

4.6 Analyses for different M

In our method, multiple individual networks are generated based on facial local regions, and the previous experiments are implemented with the number of local patches $M = 16$. Therefore, we also analyse the number

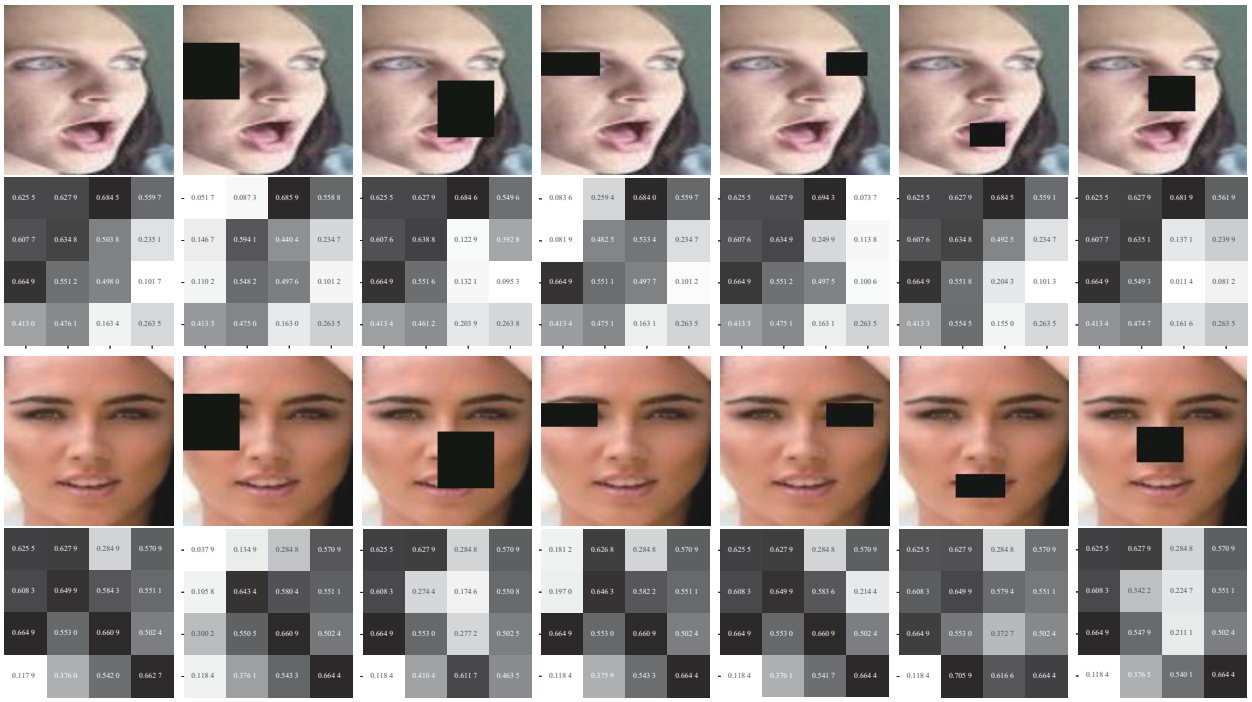


Fig. 11 Local weights of 16 patches of each face on RAF-DB[23] obtained by the proposed model. The first column shows the results corresponding to original images, and the second to the seventh columns show the results corresponding to obscured images.

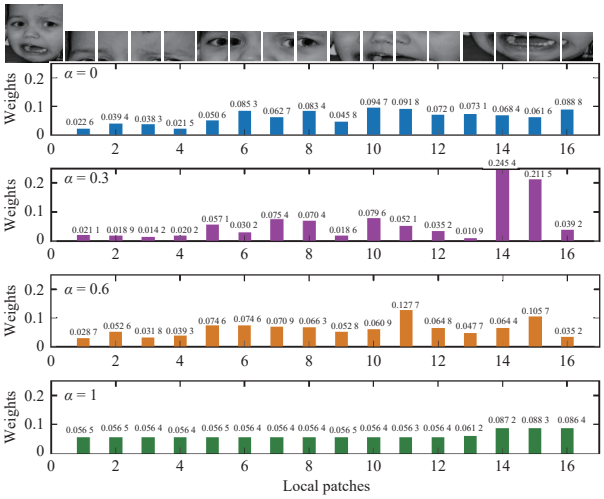


Fig. 12 Change in the non-local weight of an image from RAF-DB[23] at different α

(M) of local patches on five datasets. In this experiment, M is set as 4, 9, 16, 25 and 36. Table 3 shows the accuracy rates with different M. In this experiment, the size of the input image is 144×144, and the size of overlapping pixels between adjacent patches is approximately one third of the size of each patch, which is computed by

$$n \times P_{size} - (n-1) \times \gamma \times P_{size} = 144 \tag{5}$$

where γ is approximately $1/3$, $n^2 = M$ and P_{size} is the size of each patch. Note that the parameters of our network except M are set as the same as previous experiments.

From Table 3, it is observed that the performance with more local regions is superior to that with fewer local regions. This implies that the size of each local region is too large to attain multiple diverse local information when M is set as a small value. However, it is also noticed that the computational complexity will be increased when M is set as a high value, and thus we finally set $M = 16$ to implement most experiments.

4.7 Analyses for overlapped pixels between local regions

In the previous experiments, 1/3 of all pixels in each patch are applied as the overlapping pixels between two neighboring patches, which is a more appropriate value, since the number of pixels overlapping between the middle patch and its two sides is only 2/3, and the information of 1/3 of the pixels at the center of the patch is still retained. If a larger number of overlapping pixels is employed, such as 1/2, the middle patch will completely overlap with the patches on both sides. If a smaller number is used, such as 1/4, the number of pixels in the overlapping region will be too small to solve the problem of regional connectivity. To analyse the influence of overlapping pixels between two patches, an experiment in which the other experimental settings are the same as before is implemented based on the RAF-DB dataset, and the result is shown in Table 4.

Table 4 shows the accuracy obtained by the proposed method based on different numbers (N) of overlapping pixels. From the results, it is seen that the performance on the test set increases slowly to plateau as the number of overlapping pixels increases. This illustrates that the more overlapping pixels there are, the larger the number of network parameters. According to our analysis, the main reason is that it is easier to introduce redundant information between adjacent patches when the number of overlapping pixels is larger.

5 Conclusions

In this paper, we propose the LNLAttenNet method to effectively explore the significance of facial crucial regions in feature learning for FER, without any landmark information. In LNLAttenNet, the global information of the facial expression is utilized to construct the non-local attention network, and the local information is utilized to supervise self-information. By the joint optimization of facial non-local and local feature vectors, LNLAttenNet can adaptively enhance more crucial regions in the process of deep feature learning. Specifically, an ensemble of multiple networks corresponding to local regions is constructed to integrate the local feature with the non-local weights, which achieves interactive guidance between the facial global and local information. The experimental results also demonstrate that some local crucial regions can be effectively enhanced in feature learning by LNLAttenNet while there is no given information on landmarks in the training model. Moreover, the proposed method fo-

Table 2 Accuracy rates (%) given by the proposed method with different α

α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
RAF	84.09	85.60	85.69	**86.15**	85.59	85.33	85.17	85.23	83.74	83.54	83.02
SFEW	55.06	55.73	56.88	**57.80**	57.34	57.11	56.65	56.88	55.96	54.59	53.67
CK+	96.02	96.75	97.56	98.18	**98.30**	97.74	97.36	96.60	96.22	96.04	95.28
MMI	67.00	67.45	68.50	68.75	**68.88**	68.25	67.50	67.38	66.93	66.50	66.25
AffectNet	57.94	58.71	**59.43**	59.28	58.03	57.80	56.83	56.86	56.71	56.66	56.63

Table 3 Accuracy (%) of the proposed method with different numbers (M) of patches

M	4	9	16	25	36
RAF	84.97	85.66	**86.15**	85.53	85.63
SFEW	55.28	56.88	57.80	**58.03**	57.80
CK+	96.22	97.17	**98.18**	97.92	97.74
MMI	67.60	67.90	68.75	**68.83**	67.13
AffectNet	58.06	58.43	**59.28**	59.06	57.97

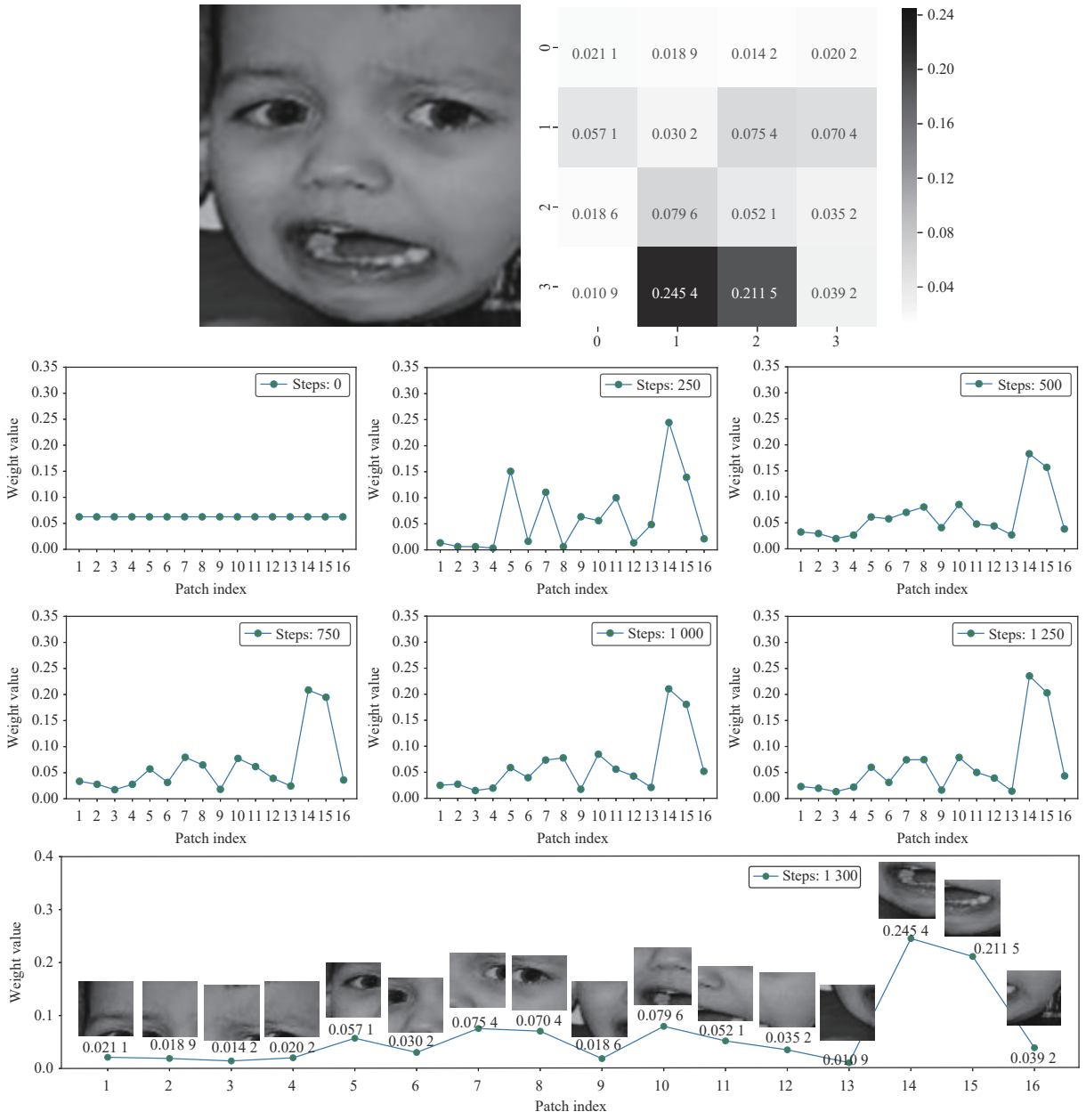
Table 4 Accuracy (%) of the proposed method with different overlapping numbers (N) of pixels

N	4	8	12	16	20	24
RAF	84.63	84.96	85.29	86.15	86.16	86.24

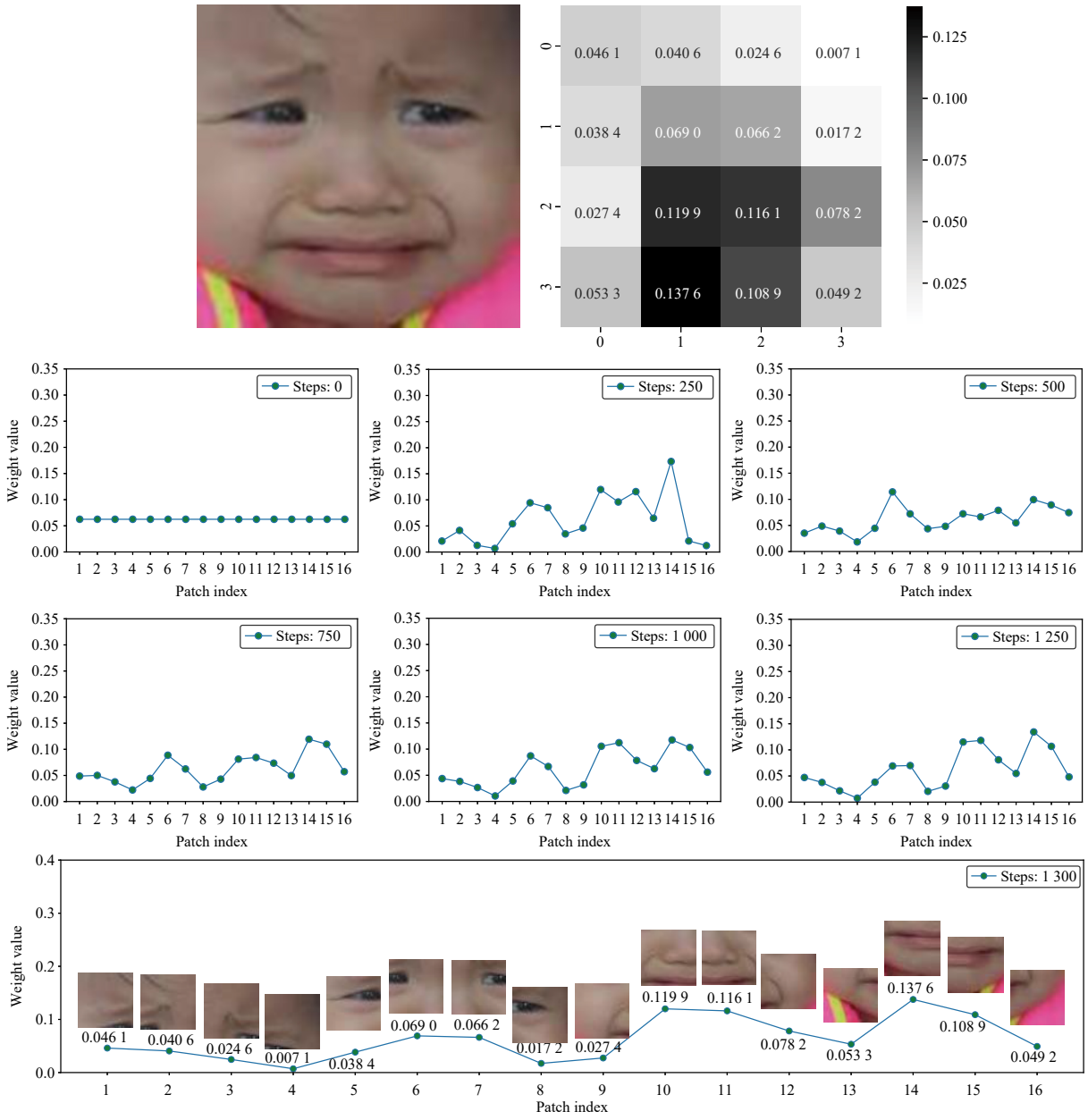
cuses on enhancing facial crucial regions in FER without any landmark information based on multiple patches, and thus we will explore it from the view of pixels for facial expressions in future works.

Appendix

Extended data. Fig. A1 shows the change of non-local weights in the training process. In Fig. A1, the first row shows the original image and its final global weights obtained by our model, and the second and third rows show the given global weights of 16 local regions in the initial,



(a) Change of non-local weights in the training process for a facial image with fear expression

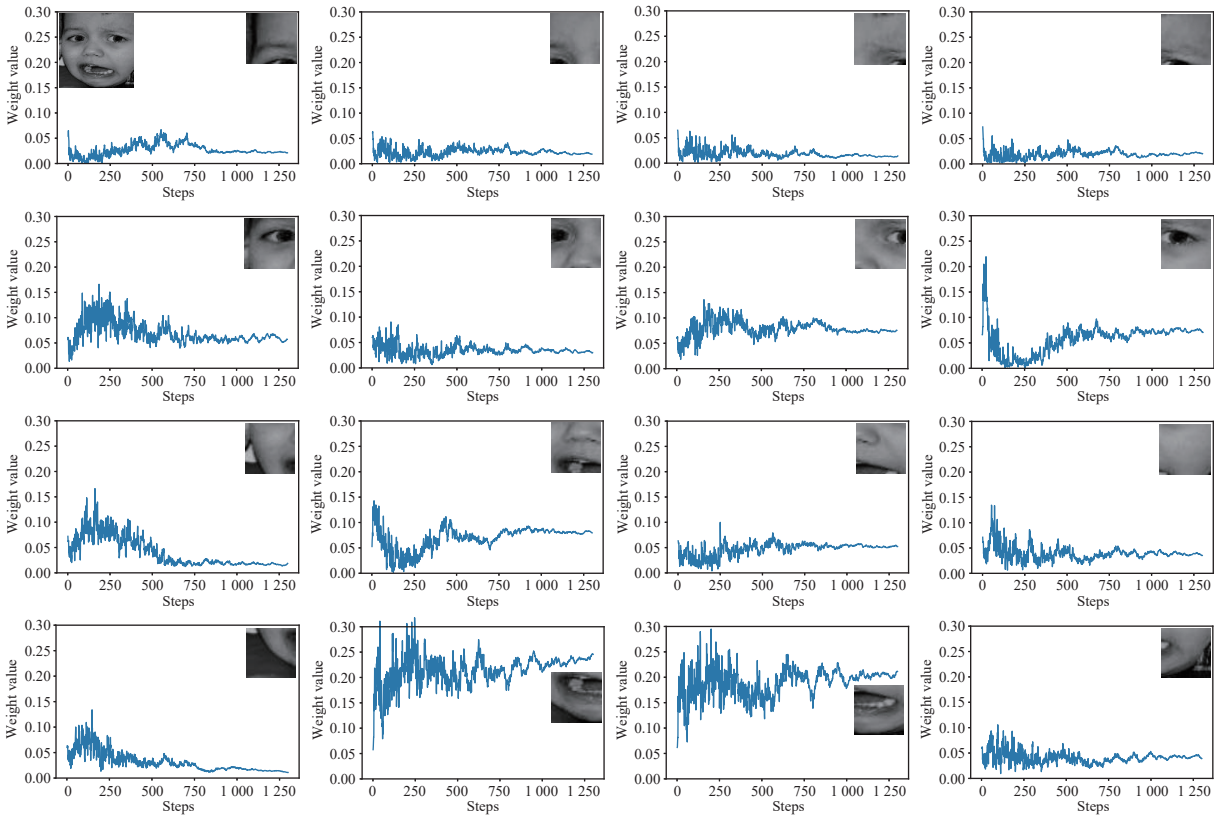


(b) Change of non-local weights in the training process for a facial image with sad expression

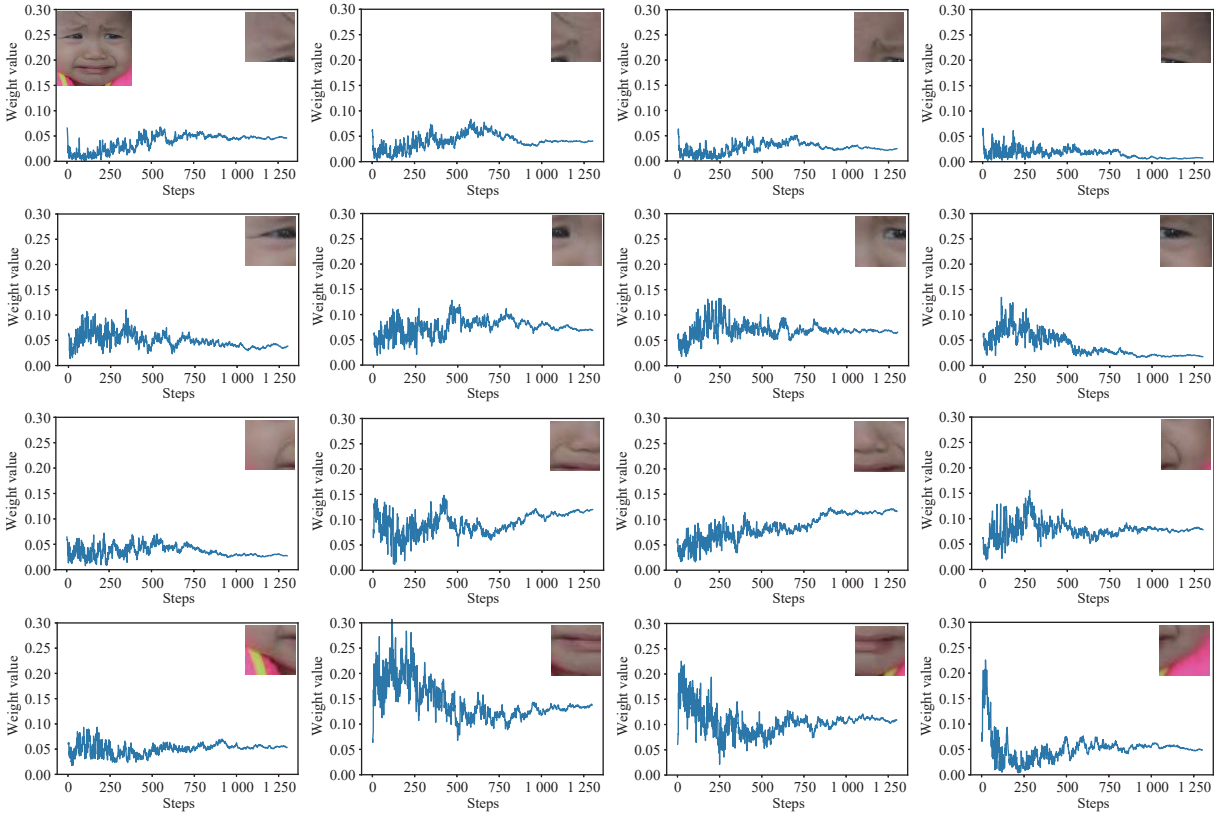
Fig. A1 16 non-local weights of two input images from RAF-DB[23]. In the first row, the input image and the non-local weights corresponding to each patch are shown. In the second and the third rows, the six figures show the non-local weights of the input images at different training stages, respectively. The last row shows the final non-local weights obtained by our model.

250th, 500th, 750th, 1 000th and 1 250th iterations, respectively, and the last row shows the final weights. From Fig. A1, it is seen that the non-local weight of each local patch is same at the beginning of training, which implies that each local region is initially regarded as having equal importance. With the training of our network, each local region is given different weights, and the higher weights are given some more discriminative regions, such as the patches (located at (4, 2) and (4, 3)), including the mouth shown in Fig. A1(a), the patches (located at (3, 2), (3, 3), (4, 2) and (4, 4)) in Fig. A1(b), etc. It illustrates that some more crucial local regions can be adaptively enhanced in the training of our network without any landmarks.

In order to better observe the change of weights, we also show the change of weights corresponding to 16 local regions in all iterations in Fig. A2. From Fig. A2, it is seen that the weight value fluctuates at the beginning of network training and it is gradually stabilized until the end of the training. Some patches that are visually more discriminative are given higher weights and some patches located at the noncrucial regions cut down with smaller weights. In summary, the analyses for non-local weights demonstrate that the proposed method can effectively



(a) Change of weights corresponding to 16 local regions in all iterations for a facial image with fear expression



(b) Change of weights corresponding to 16 local regions in all iterations for a facial image with sad expression

Fig. A2 Change of weights $\boldsymbol{w}^g$ corresponding to 16 local regions of two images from RAF-DB[23] in the training process of LNLAttenNet. The abscissa represents the number of iterations in the training process and the ordinate represents the magnitude of the weight corresponding to each iteration.

automatically enhance the significance of facial crucial regions in deep feature learning, without any given prior information of facial crucial regions.

Declarations of conflict of interest

The authors declared that they have no conflicts of interest to this work.

Open Access

References

[1] C. Darwin, P. Ekman. *The Expression of the Emotions in Man and Animals*, 3rd ed., Oxford, UK: Oxford University Press, 1998.

[2] R. Buck, R. E. Miller, W. F. Caul. Sex, personality, and physiological variables in the communication of affect via facial expression. *Journal of Personality and Social Psychology*, vol. 30, no. 4, pp. 587–596, 1974. DOI: 10.1037/h0037041.

[3] M. C. Smith, M. K. Smith, H. Ellgring. Spontaneous and posed facial expression in parkinson′s disease. *Journal of the International Neuropsychological Society*, vol. 2, no. 5, pp. 383–391, 1996. DOI: 10.1017/S1355617700001454.

[4] C. A. Corneanu, M. O. Simón, J. F. Cohn, S. E. Guerrero. Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016. DOI: 10.1109/TPAMI.2016.2515606.

[5] A. Majumder, L. Behera, V. K. Subramanian. Automatic facial expression recognition system using deep network-based data fusion. *IEEE Transactions on Cybernetics*, vol. 48, no. 1, pp. 103–114, 2018. DOI: 10.1109/TCYB.2016.2625419.

[6] W. C. Xie, L. L. Shen, J. M. Duan. Adaptive weighting of handcrafted feature losses for facial expression recognition. *IEEE Transactions on Cybernetics*, vol. 51, no. 5, pp. 2787–2800, 2021. DOI: 10.1109/TCYB.2019.2925095.

[7] E. L. Rosenberg, P. Ekman. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS)*, 3rd ed., Oxford, UK: Oxford University Press, 2020.

[8] S. F. Wang, G. Z. Peng, S. Y. Chen, Q. Ji. Weakly supervised facial action unit recognition with domain knowledge. *IEEE Transactions on Cybernetics*, vol. 48, no. 11, pp. 3265–3276, 2018. DOI: 10.1109/TCYB.2018.2868194.

[9] H. K. Ekenel, R. Stiefelhagen. Why is facial occlusion a challenging problem? In *Proceedings of the 3rd International Conference on Biometrics*, Springer, Alghero, Italy, pp. 299–308, 2009. DOI: 10.1007/978-3-642-01793-3_31.

[10] L. Zhong, Q. S. Liu, P. Yang, J. Z. Huang, D. N. Metaxas. Learning multiscale active facial patches for expression analysis. *IEEE Transactions on Cybernetics*, vol. 45, no. 8, pp. 1499–1510, 2015. DOI: 10.1109/TCYB.2014.2354351.

[11] Y. R. Fan, J. C. K. Lam, V. O. K. Li. Multi-region ensemble convolutional neural network for facial expression recognition. In *Proceedings of the 27th International Conference on Artificial Neural Networks*, Springer, Rhodes, Greece, pp. 84–94, 2018. DOI: 10.1007/978-3-030-01418-6_9.

[12] Y. Li, J. B. Zeng, S. G. Shan, X. L. Chen. Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2019. DOI: 10.1109/TIP.2018.2886767.

[13] S. L. Happy A. Routray. Automatic facial expression recognition using features of salient facial patches. *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 1–12, 2015. DOI: 10.1109/TAFFC.2014.2386334.

[14] K. Wang, X. J. Peng, J. F. Yang, D. B. Meng, Y. Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020. DOI: 10.1109/TIP.2019.2956143.

[15] M. Y. Liu, S. G. Shan, R. P. Wang, X. L. Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, USA, pp. 1749–1756, 2014. DOI: 10.1109/CVPR.2014.226.

[16] C. F. Shan, S. G. Gong, P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009. DOI: 10.1016/j.imavis.2008.08.005.

[17] I. Kotsia, I. Pitas. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 172–187, 2007. DOI: 10.1109/TIP.2006.884954.

[18] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, San Francisco, USA, pp. 94–101, 2010. DOI: 10.1109/CVPRW.2010.5543262.

[19] M. Pantic, M. Valstar, R. Rademaker, L. Maat. Web-based database for facial expression analysis. In *Proceedings of IEEE International Conference on Multimedia and Expo*, Amsterdam, The Netherlands, pp. 317–321, 2005. DOI: 10.1109/ICME.2005.1521424.

[20] M. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings of*

the 3rd IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, pp. 200–205, 1998. DOI: 10.1109/AFGR.1998.670949.

[21] G. Y. Zhao, X. H. Huang, M. Taini, S. Z. Li, M. Pietikäinen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011. DOI: 10.1016/j.imavis.2011.07.002.

[22] S. Li, W. H. Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, 2022. DOI: 10.1109/TAFFC.2020.2981446.

[23] S. Li, W. H. Deng, J. P. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 2584–2593, 2017. DOI: 10.1109/CVPR.2017.277.

[24] A. Mollahosseini, B. Hasani, M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2019. DOI: 10.1109/TAFFC.2017.2740923.

[25] C. F. Benitez-Quiroz, R. Srinivasan, A. M. Martinez. EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 5562–5570, 2016. DOI: 10.1109/CVPR.2016.600.

[26] L. T. Mou, C. Zhou, P. T. Xie, P. F. Zhao, R. Jain, W. Gao, B. C. Yin. Isotropic self-supervised learning for driver drowsiness detection with attention-based multimodal fusion. *IEEE Transactions on Multimedia*, vol. 25, pp. 529–542, 2021. DOI: 10.1109/TMM.2021.3128738.

[27] L. T. Mou, C. Zhou, P. F. Zhao, B. Nakisa, M. N. Rastgoo, R. Jain, W. Gao. Driver stress detection via multimodal fusion using attention-based CNN-LSTM. *Expert Systems with Applications*, vol. 173, Article number 114693, 2021. DOI: 10.1016/j.eswa.2021.114693.

[28] L. Song, J. F. Yang, Q. Z. Shang, M. A. Li. Dense face network: A dense face detector based on global context and visual attention mechanism. *Machine Intelligence Research*, vol. 19, no. 3, pp. 247–256, 2022. DOI: 10.1007/s11633-022-1327-2.

[29] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition, [Online], Available: https://arxiv.org/abs/1409.1556, 2014.

[30] C. Szegedy, W. Liu, Y. Q. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich. Going deeper with convolutions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 1–9, 2015. DOI: 10.1109/CVPR.2015.7298594.

[31] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 770–778, 2016. DOI: 10.1109/CVPR.2016.90.

[32] P. Hu, D. Q. Cai, S. D. Wang, A. B. Yao, Y. R. Chen. Learning supervised scoring ensemble for emotion recognition in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, Glasgow, UK, pp. 553–560, 2017. DOI: 10.1145/3136755.3143009.

[33] D. Acharya, Z. W. Huang, D. P. Paudel, L. van Gool. Covariance pooling for facial expression recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, Salt Lake City, USA, pp. 480–4807, 2018. DOI: 10.1109/CVPRW.2018.00077.

[34] S. Li, W. H. Deng. Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. *International Journal of Computer Vision*, vol. 127, no. 6–7, pp. 884–906, 2019. DOI: 10.1007/s11263-018-1131-1.

[35] H. Y. Yang, L. J. Yin. CNN based 3D facial expression recognition using masking and landmark features. In *Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction*, IEEE, San Antonio, USA, pp. 556–560, 2017. DOI: 10.1109/ACII.2017.8273654.

[36] W. Q. Wu, Y. J. Yin, Y. Y. Wang, X. G. Wang, D. Xu. Facial expression recognition for different pose faces based on special landmark detection. In *Proceedings of the 24th International Conference on Pattern Recognition*, IEEE, Beijing, China, pp. 1524–1529, 2018. DOI: 10.1109/ICPR.2018.8545725.

[37] J. B. Zeng, S. G. Shan, X. L. Chen. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 227–243, 2018. DOI: 10.1007/978-3-030-01261-8_14.

[38] Y. L. Gan, J. Y. Chen, L. H. Xu. Facial expression recognition boosted by soft label with a diverse ensemble. *Pattern Recognition Letters*, vol. 125, pp. 105–112, 2019. DOI: 10.1016/j.patrec.2019.04.002.

[39] H. Y. Yang, U. Ciftci, L. J. Yin. Facial expression recognition by de-expression residue learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 2168–2177, 2018. DOI: 10.1109/CVPR.2018.00231.

[40] F. F. Zhang, T. Z. Zhang, Q. R. Mao, C. S. Xu. Joint pose and expression modeling for facial expression recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 3359–3368, 2018. DOI: 10.1109/CVPR.2018.00354.

[41] S. C. Zhao, C. Lin, P. F. Xu, S. D. Zhao, Y. C. Guo, R. Krishna, G. G. Ding, K. Keutzer. CycleemotionGAN: Emotional semantic consistency preserved cycleGAN for adapting image emotions. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, Honolulu, USA, pp. 2620–2627, 2019. DOI: 10.1609/aaai.v33i01.33012620.

[42] L. J. Fan, W. B. Huang, C. Gan, J. Z. Huang, B. Q. Gong. Controllable image-to-video translation: A case study on facial expression generation. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, Honolulu, USA, pp. 3510–3517, 2019. DOI: 10.1609/aaai.v33i01.33013510.

[43] R. L. Wu, G. J. Zhang, S. J. Lu, T. Chen. Cascade EF-GAN: Progressive facial expression editing with local focuses. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 5020–5029, 2020. DOI: 10.1109/CVPR42600.2020.00507.

[44] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, F. Moreno-Noguer. GANimation: Anatomically-aware facial animation from a single image. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 835–851, 2018. DOI: 10.1007/978-3-

030-01249-6_50.

[45] Y. Choi, M. Choi, M. Kim, J. W. Ha, S. Kim, J. Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 8789–8797, 2018. DOI: 10.1109/CVPR.2018.00916.

[46] P. Barros, G. I. Parisi, C. Weber, S. Wermter. Emotion-modulated attention improves expression recognition: A deep learning model. *Neurocomputing*, vol. 253, pp. 104–114, 2017. DOI: 10.1016/j.neucom.2017.01.096.

[47] X. H. Wang, M. Z. Peng, L. J. Pan, M. Hu, C. H. Jin, F. J. Ren. Two-level attention with two-stage multi-task learning for facial emotion recognition. *Journal of Visual Communication and Image Representation*, vol. 62, pp. 217–225, 2019. DOI: 10.1016/j.jvcir.2019.05.009.

[48] O. Ronneberger, P. Fischer, T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the 18th International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Munich, Germany, pp. 234–241, 2015. DOI: 10.1007/978-3-319-24574-4_28.

[49] T. Falk, D. Mai, R. Bensch, Ö. Çiçek, A. Abdulkadir, Y. Marrakchi, A. Böhm, J. Deubner, Z. Jäckel, K. Seiwald, A. Dovzhenko, O. Tietz, C. Dal Bosco, S. Walsh, D. Saltukoglu, T. L. Tay, M. Prinz, K. Palme, M. Simons, I. Diester, T. Brox, O. Ronneberger. U-Net: Deep learning for cell counting, detection, and morphometry. *Nature Methods*, vol. 16, no. 1, pp. 67–70, 2019. DOI: 10.1038/s41592-018-0261-2.

[50] Z. X. Zhang, Q. J. Liu, Y. H. Wang. Road extraction by deep residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018. DOI: 10.1109/LGRS.2018.2802944.

[51] T. Y. Lin, P. Dollár, R. Girshick, K. M. He, B. Hariharan, S. Belongie. Feature pyramid networks for object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 936–944, 2017. DOI: 10.1109/CVPR.2017.106.

[52] J. Long, E. Shelhamer, T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 3431–3440, 2015. DOI: 10.1109/CVPR.2015.7298965.

[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, USA, pp. 6000–6010, 2017. DOI: 10.5555/3295222.3295349.

[54] X. L. Wang, R. Girshick, A. Gupta, K. M. He. Non-local neural networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 7794–7803, 2018. DOI: 10.1109/CVPR.2018.00813.

[55] H. S. Zhao, J. P. Shi, X. J. Qi, X. G. Wang, J. Y. Jia. Pyramid scene parsing network. In *Proceed-ings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 6230–6239, 2017. DOI: 10.1109/CVPR.2017.660.

[56] A. Dhall, R. Goecke, S. Lucey, T. Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *Proceedings of IEEE International Conference on Computer Vision Workshops*, Barcelona, Spain, pp. 2106–2112, 2011. DOI: 10.1109/ICCVW.2011.6130508.

[57] J. Goldberger, E. Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France, 2017.

[58] Y. D. Wen, K. P. Zhang, Z. F. Li, Y. Qiao. A discriminative feature learning approach for deep face recognition. In *Proceedings of the 14th European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, pp. 499–515, 2016. DOI: 10.1007/978-3-319-46478-7_31.

[59] S. K. Chen, J. F. Wang, Y. D. Chen, Z. C. Shi, X. Geng, Y. Rui. Label distribution learning on auxiliary label space graphs for facial expression recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 13981–13990, 2020. DOI: 10.1109/CVPR42600.2020.01400.

[60] A. P. Fard, M. H. Mahoor. Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild. *IEEE Access*, vol. 10, pp. 26756–26768, 2022. DOI: 10.1109/ACCESS.2022.3156598.

Guanghui Shi received the B. Sc. degree in electronics and information engineering from Wuhan University of Technology, China in 2018, and the M. Sc. degree in electronics and communication engineering from Xidian University, China in 2021. He is currently a Ph. D. degree candidate in computer science and technology at Xidian University, China.

His research interests include deep learning, facial expression recognition and visual pattern mining.

E-mail: ghshi@stu.xidian.edu.cn

ORCID iD: 0000-0003-2230-207X

Shasha Mao received the Ph. D. degree in circuit and system from Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University, China in 2014. She is currently an associate professor at Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, School of Artificial Intelligence, Xidian University, China. She worked as a research fellow in Nanyang Technological University and Singapore University of Technology and Design, Singapore from 2014 to 2018.

Her research interests include ensemble learning, deep learning, imbalanced learning, facial expression recognition and SAR image registration.

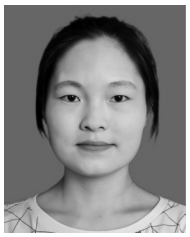
E-mail: ssmao@xidian.edu.cn (Corresponding author)

ORCID iD: 0000-0003-3308-1794

Shuiping Gou received the B. Sc. and M. Sc. degrees in computer science and technology from Xidian University, China in 2000 and 2003, respectively, and the Ph. D. degree in pattern recognition and intelligent system from Xidian University, China in 2008. She is currently a professor with Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, School of Artificial Intelligence, Xidian University, China.

Her research interests include machine learning, data mining, remote sensing image analysis and medical image analysis.

E-mail: shpgou@mail.xidian.edu.cn
ORCID iD: 0000-0002-2619-6481



Dandan Yan received the B. Sc. degree in computer science and technology from Xi′an University of Technology, China in 2021. She is currently a master student in computer science and technology at School of Artificial Intelligence, Xidian University, China.

Her research interests include deep learning, facial expression recognition and label distribution learning.

E-mail: ddanyan0319@stu.xidian.edu.cn
ORCID iD: 0000-0002-9195-2618

Licheng Jiao received the B. Sc. degree in electronic engineering from Shanghai Jiao Tong University, China in 1982, the M. Sc. and Ph. D. degrees in electronic engineering from Xi′an Jiaotong University, China, in 1984 and 1990, respectively. From 1990 to 1991, he was a postdoctoral fellow with National Key Laboratory for Radar Signal Processing, Xidian University, China. Since 1992, he has been a professor with School of Electronic Engineering, Xidian University, China. Currently, he is a professor with School of Artificial Intelligence, Xidian University, China, and he is also the director of Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education of China, Xidian University, China. He is in charge of approximately 40 important Scientific research projects. He has authored or coauthored more than 20 monographs and 100 papers in international journals and conferences. He is a member of the IEEE Xi′an Section Execution Committee, the Chairman of Awards and Recognition Committee, the Vice Board Chairperson of the Chinese Association of Artificial Intelligence, the Councilor of the Chinese Institute of Electronics, the Committee Member of the Chinese Committee of Neural Networks, and an Expert of Academic Degrees Dommittee of the State Council.

His research interests include image processing, natural computation, machine learning and intelligent information processing.

E-mail: lchjiao@mail.xidian.edu.cn
ORCID iD: 0000-0003-3354-9617

Lin Xiong received the Ph. D. degree in pattern recognition & intelligent system from Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University, China in 2015. Currently, he works as a senior algorithm expert in SenseTime, China. He was a research scientist at JD Finance America Corporation from 2018 to 2022.

His research interests include neural implicit rendering, neural radiance field, distributed model parallelism, unconstrained/large-scale face recognition, deep learning architecture engineering, person reidentification, face recognition, Riemannian manifold optimization, sparse and low-rank matrix factorization.

E-mail: xionglin@sensetime.com
ORCID iD: 0000-0003-3545-227X