IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 17, 2019, accepted December 31, 2019, date of publication January 3, 2020, date of current version January 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2963913

Multiple Attention Network for Facial Expression Recognition

YANLING GAN[1], JINGYING CHEN[1], ZONGKAI YANG[1], AND LUHUI XU[2]
[1]National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China
[2]Department of Computer Science, Guangxi Normal University, Guilin 541004, China

Corresponding authors: Jingying Chen (chenjy@mail.ccnu.edu.cn) and Zongkai Yang (zkyang@mail.ccnu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1004504, in part by the National Natural Science Foundation under Grant 61977027 and Grant 61772380, in part by the Hubei Province Technological Innovation Major Project under Grant 2019AAA044, in part by the Science and Technology Major Project of Hubei Province (Next-Generation AI Technologies) under Grant 2019AEA170, in part by the Foundation for Innovative Research Groups of Hubei Province under Grant 2017CFA007, and in part by the Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE under Grant CCNU19Z02002, Grant CCNU18KFY02, and Grant 2019CXZZ014.

ABSTRACT One key challenge in facial expression recognition (FER) is the extraction of discriminative features from critical facial regions. Because of their promising ability to learn discriminative features, visual attention mechanisms are increasingly used to address pattern recognition problems. This paper presents a novel multiple attention network that simulates humans' coarse-to-fine visual attention to improve expression recognition performance. In the proposed network, a region-aware sub-net (RASnet) learns binary masks for locating expression-related critical regions with coarse-to-fine granularity levels and an expression recognition sub-net (ERSnet) with a multiple attention (MA) block learns comprehensive discriminative features. Embedded in the convolutional layers, the MA block fuses diversified attention using the learned masks from the RASnet. The MA block contains a hybrid attention branch with a series of sub-branches, where each sub-branch provides region-specific attention. To explore the complementary benefits of diversified attention, the MA block also has a weight learning branch that adaptively learns the contributions of the different critical regions. Experiments have been carried out on two publicly available databases, RAF and CK+, and the reported accuracies are 85.69% and 96.28%, respectively. The results indicate that our method achieves competitive or better performance than state-of-the-art methods.

INDEX TERMS Facial expression recognition, multiple attention network, binary masks.

I. INTRODUCTION

Expression, a common form of nonverbal communication, conveys important cues for emotional states and intentions. Automatic facial expression recognition (FER) has many practical applications, such as in improving human-computer interaction and remote education [1]–[4]. However, irrelevant facial information (e.g., hair and hat) and complex background clutter create problems when using automatic FER. In contrast, human observers can pay selective attention to the expression-related parts of a facial image while screening out the irrelevant components, resulting in high-level FER performance. Motivated by the attention mechanism, many methods have been developed to improve how FER models distinguish between different expressions and eliminate or suppress irrelevant information.

For FER task, traditional methods use detection techniques to reduce the negative effects of irrelevant information. For example, some methods that detected expression-related facial components (e.g., AUs [5]) and small patches of interest (e.g., eyes, nose, and mouth [6], [7]) aimed to extract accurate features from critical facial regions. Several recent FER studies [8]–[10] used deep networks to mimic the attention mechanism and achieved excellent FER performance. However, two of these studies [8], [9] simply adopted single-level (i.e., global-level) attention without any consideration for diversified saliencies, which may distract attention to expression-irrelevant components. Li *et al.* [10] adopted region-level attention to examine the importance of different regions, but this method cannot learn discriminative features.

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin.

Therefore, there is a pressing need for a FER method with the following properties:

1. It should automatically locate critical facial regions, thereby eliminating the influence of irrelevant facial parts.
2. It should fuse diversified attention to effectively learn discriminative features.

Therefore, this paper presents a novel multiple attention network that mimics coarse-to-fine attention in humans to improve FER performance. The proposed network consists of two sub-nets. The first is a region-aware sub-net (RASnet) that learns masks to automatically locate critical regions with different scales and positions, and it is trained using ground-truth binary masks generated with relatively few landmarks. The second is an expression recognition sub-net (ERSnet) that learns discriminative features, and it is supported by a multiple attention (MA) block built with the learned masks. Specifically, the MA block contains a hybrid attention branch with a series of sub-branches, and each sub-branch specializes attention for one region by combing region attention and channel attention. The former aims to locate critical regions, and the latter to learn region-specific discriminative features. In addition, to aid in feature discrimination, the MA block also has a weight learning branch, which learns weight vectors and then adaptively fuses the learned features from different regions.

The main contributions of our work are as follows:

1. A novel multiple attention network that mimics coarse-to-fine visual attention to learn discriminative features from expression-related regions is proposed to improve FER performance.
2. The proposed framework includes two sub-nets: a RASnet to automatically locate critical regions and an ERSnet to learn discriminative features from these critical regions.
3. A MA block is included to address region-specific attentions and fuse various attentions. The MA block is embedded in the convolutional layers to help the ERSnet focus on learning discriminative features from expression-related regions.
4. Experiments were conducted on two databases to demonstrate that the proposed approach achieves competitive performance compared to state-of-the-art methods.

II. RELATED WORK

Our method shares several similarities with other FER methods and deep attention models. In general, deep learning models have shown promising results, which inspired us to mimic the attention mechanism in a deep learning framework to achieve better FER performance.

A. FACIAL EXPRESSION RECOGNITION METHODS

FER is a widely studied topic in the pattern recognition field because it has a wide range of practical applications. A key component of a FER system is the extraction of expression feature. Because facial images often present diverse backgrounds and facial attributes, accurate discriminative feature extraction is crucial to optimizing FER performance.

To boost FER performance, many methods focus on critical facial parts, such as the eyes, nose, and mouth. One method [7] divided each facial image into several local patches and then explored the common patches shared by all expressions and the expression-specific ones using multi-task sparse learning to extract expression features. Similarly, Happy et al. [6] detected salient facial patches, and then extracted and concatenated features from these patches for training classifiers.

Many studies focused on improving facial representation. To learn mid-level representations, Liu *et al.* [11] modeled video images as spatial-temporal manifolds and then aligned them using a universal manifold model. Compared to traditional methods, deep learning ones – such as convolutional neural network (CNN) and recurrent neural network (RNN) – have more potential in feature learning and classification due to their end-to-end learning capabilities, and this is supported by many pattern recognition tasks that have included expression recognition [10], [12]–[15].

To improve expression discrimination, Li *et al.* [14] proposed a deep locality-preserving CNN (DLP-CNN), in which locality-preserving loss was developed to keep locality closeness. For the same purpose, Cai *et al.* [16] proposed island loss that can minimize the intra-class distances of deep features while maximizing their inter-class distances. Liu et al. [17] jointly optimized (N+M)-tuplet cluster loss and softmax loss in their FER framework, and they proposed using an identity-aware hard-negative mining strategy to achieve an identity-invariant property.

Jung *et al.* [18] presented a joint fine-tuning strategy with a deep framework. This framework contained two different CNNs: one for extracting temporal appearance features from video sequences and another for extracting temporal geometry features from landmark trajectories. The outputs of these two networks were combined using weighted summation, and the whole framework was trained via a joint fine-tuning method.

In literature [19], the authors constructed a deeper FER network by utilizing inception layers, and they aimed to expand the depth and width of the network without increasing its computational cost. Fan *et al.* [20] learned deep features from more crucial components. They cropped different local facial regions, and then used the paired images (i.e., the cropped region and the whole image) to train multiple CNN classifiers for ensemble. Zhao *et al.* [21] embedded a feature selection mechanism into an advanced deep architecture to filter irrelevant features and emphasize correlated features. Liu *et al.* [15] proposed an AU-aware deep networks (AUDN), and they used a feature selection method in the middle layer to select AU-aware receptive fields to learn expression features.

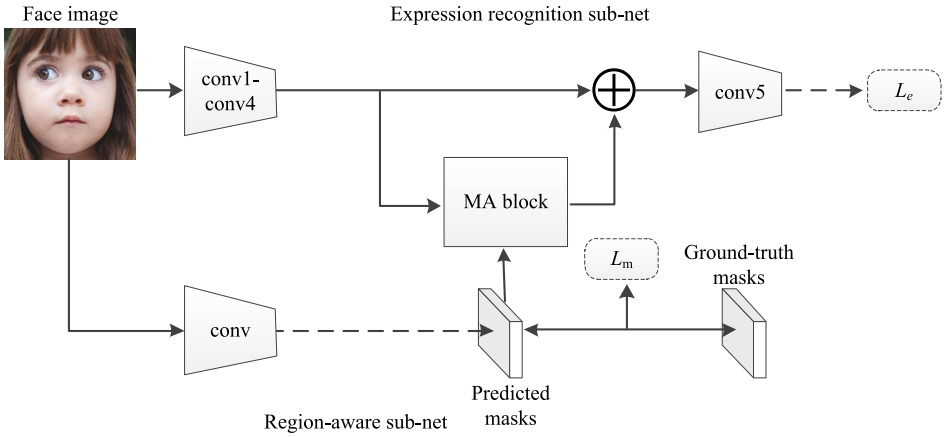

FIGURE 1. An overview of the proposed framework, which contains two sub-nets. The first one is a region-aware sub-net that learns various region masks automatically, and the architecture and configuration are shown in table 1. The second one is an expression recognition sub-net containing a backbone network and a MA block, and the configuration is shown in figure 2.

B. DEEP ATTENTION MODEL

When people observe objects, they tend to focus selectively on the critical parts rather than the whole object to obtain important clues. Inspired by this, some visual attention models adaptively emphasize the important information while suppressing the irrelevant information, and they have achieved impressive results for addressing problems in pattern recognition, computer vision, and other fields [22]–[25].

Several recent FER studies [8]–[10] have mimicked the attention mechanism with a deep learning framework. Sun *et al.* [9] integrated visual attention into CNN to learn features of interest. However, their method learn only global attention from the whole images, thus is prone to distract attention to irrelevant components. This may lead to negative outcomes for expression recognition methods.

Jang *et al.* [8] mimicked human visual fixation over time to generate sequences for still-images, and they extracted the CNN features of these sequences to train an RNN for recognizing facial attribute, including expression. However, this method customize attention in fixed positions, which may lead to poor generalization since these positions do not consider the differences of expressions and individuals. Notably, different expressions reflect different facial feature movements, showing different saliencies. In addition, different individuals may express the same expression in different ways. For example, when expressing happiness, one person may smile with an open mouth, while another may smile with a closed mouth. Furthermore, mimicking attention in input [8] may limit the representation learning capability of the deep network, because low-level semantics are messy. Consequently, early attention inputs are prone to introducing irrelevant information.

Li *et al.* [10] proposed Patch-Gated Convolution Neutral Network (PG-CNN) for FER. They used landmark information to extract small patches of interest, and embedded PG-Units in their network to learn the weights for these patches to obtain region-level attention, enabling their method to be occlusion-aware. However, region-level attention can not learn more discriminative features.

TABLE 1. Configuration parameters in the RASnet.

Layer	1	2	3	4	5	6	7	8	9	10
	conv+ ReLU	MPool	conv+ ReLU	MPool	conv+ ReLU	MPool	fc	fc	reshape	sigmoid
Kernel	5	2	5	2	3	2	-	-	-	-
Output	64	-	128	-	128	-	2048	1056	-	-
Stride	3	2	3	2	1	2	-	-	-	-
Pad	1	-	1	-	0	-	-	-	-	-
Shape	-	-	-	-	-	-	-	-	8×14×14	-

III. METHODOLOGY

To learn discriminative features, our CNN mimics visual attention that accounts for the diversity of expressions and individuals. Unlike other deep attention models, our model fuses diversified attention that covers different granularities, which improves its ability to learn comprehensive discriminative features. Notably, fusing diversified attention helps the model to avoid attention distraction and to focus on only the most important facial information.

The overview of the proposed framework is shown in figure 1. It consists of a RASnet and an ERSnet. The RASnet learns binary masks to automatically locate critical regions, and the ERSnet performs feature learning and expression recognition using the learned masks.

In pattern recognition tasks, full convolution networks are often tasked with learning binary masks. One benefit is that full convolution network can retain spatial information. But when learning multiple binary masks where overlap exist, a network may cause competitions between different masks since spatial information is retained. To avoid this, we regress the masks using fully connected (fc) layers at the bottom of RASnet. The configuration parameters of RASnet are shown in table 1. It contains three convolution (conv) layers and two

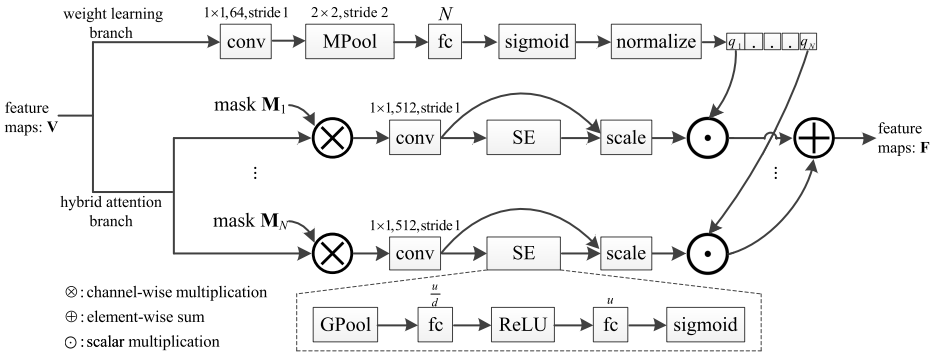

FIGURE 2. A visualization of the MA block. It contains a hybrid attention branch and a weight learning branch. The parameter settings are shown in the figure.

fc layers. Each conv layer is followed by a ReLu and a MPool. The last fc layer is followed by a reshape layer and a sigmoid layer.

The ERSnet contains two main components: a backbone and a MA block. The backbone adopts VGG16 architecture (including five conv blocks and three fc layers), but the number of the neuron in the last fc layer is changed to the expression class number. The MA block is embedded between the fourth and fifth conv blocks of the backbone to benefit from two aspects. First, the network has learnt deep features with high-level semantics at the fourth block, and the embedded MA block is consequently able to learn more accurate features. Second, the outputted feature maps from the fourth block have an appropriate scale, which makes it appropriate to implement multiple attention at this location.

The MA block is designed with consideration for three important factors. First, the network should learn features from critical regions while filtering out as many irrelevant facial features as possible. Second, the network should mimic coarse-to-fine attention to fuse diversified saliencies. Third, the learned salient features should be fused adaptively. For differently categorized expressions, the discriminative features should show significant differences. For expressions in the same category, individual differences should be minimized.

To address these factors, the MA block contains two branches designed to fuse diversified attention: a hybrid attention branch and a weight learning branch, as shown in figure 2. The hybrid attention branch contains N sub-branches, and each sub-branch learns region-specific attention for its current facial part. Each sub-branch starts with a region attention layer, which inputs the learned region masks from the RASnet and the shared convolutional features from the fourth conv block of ERSnet. The following is a 1×1 convolution layer that serves as a buffer, aiming to learn diversified representations for different regions. Next, a Squeeze-and-Excitation (SE) [29] block is implemented to learn region-specific channel attention. Meanwhile, the weight learning branch learns the weights for different regions to adaptively fuse diversified discriminative features. It contains five layers, and the first two layers reduce the numbers of dimension and channel. The following layer

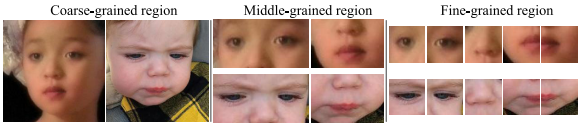

FIGURE 3. Example images of facial regions with different granularities. From left to right are coarse-grained regions (i.e., the whole face), middle-grained regions (i.e., eyes together with brows and mouth together with nose), and fine-grained regions (i.e., left eye, right eye, left mouth corner, and right mouth corner), respectively.

is a fc layer that transforms the input into a N-dimensional weight vector. And the last two layers transform the weight vector to a normalized probability vector.

A. REGION-AWARE SUB-NET

The RASnet learns masks for multiple critical regions. To train the FER classifier, irrelevant facial parts are filtered out using the learned masks. Taking an image I as input, the RASNet predicts mask $\mathbf{M}^{N\times A\times B} = [\mathbf{M}_1, \ldots, \mathbf{M}_r, \ldots, \mathbf{M}_N]$ for N regions, which can be denoted as

$$\mathbf{M} = \sigma(\phi(g(\Omega : I))) \tag{1}$$

where $g(\cdot)$ represents the convolutional and full connected operations, and Ω represents all the parameters. $g(\cdot)$ outputs a vector with length $N \times A \times B$. ϕ denotes reshape operation, and it produces N maps with size $A \times B$, which is the same with the inputted conv features from the MA block. σ is the sigmoid function.

We obtain $\mathbf{M}$ by training the parameters of RASnet via Euclidean loss. For one sample, the loss is computed by formula (2):

$$L_m = \sum_r \sum_{i,j} (M_r(i,j) - \hat{M}_r(i,j))^2 \tag{2}$$

where $\hat{\mathbf{M}} = [\hat{\mathbf{M}}_1, \ldots, \hat{\mathbf{M}}_r, \ldots, \hat{\mathbf{M}}_N]$ is the ground-truth mask, and (i, j) is the element coordinate. For one batch, the total loss is computed as follows:

$$L_M = \frac{1}{K} \sum_{k=1}^{K} L_m^k \tag{3}$$

FIGURE 4. The eight selected facial landmarks as indicated on facial images from RAF [14].

TABLE 2. The region settings.

Region r	(x, y)	(w, h)
whole image	$(W/2, H/2)$	(W, H)
eyes together with brows	$(l_2 + l_5)/2$	$(0.36W, 0.72H)$
mouth together with nose	$(l_1 + l_8)/2$	$(0.36W, 0.54H)$
left eye right eye nose, left mouth corner and right mouth corner	$(l_2 + l_3)/2$ $(l_4 + l_5)/2$ The centers are set as l_1, l_6 and l_7, respectively.	$(0.27W, 0.27H)$

where K is the batch size, and L_m^k is the loss for the k-th sample.

$\hat{\mathbf{M}}$ is generated using the following method.

1) GROUND-TRUTH MASK GENERATION

As suggested in some FER literatures [6], [7], the regions around nose, eyes, and mouth are extremely useful for learning discriminative expression representations. For one image, we select $N = 8$ regions (i.e., whole image, eyes together with brows, mouth together with nose, left eye, right eye, nose, left mouth corner, and right mouth corner), as shown in figure 3. These regions provide coarse-to-fine granularities. The whole face provides a coarse global view of the face. Eyes together with brows provides an overall view of the eyes, while left eye and right eye respectively provide different local views. The mouth together with nose provides an overall view of the middle and lower facial part, while the nose, left mouth corner and right mouth corner provide diversified local views for their respective facial parts.

The proposed method learns discriminative features by fusing diversified attention that target these eight expression-related regions. Eight facial landmarks (as shown in figure 4) are used to generate ground-truth masks that guide the learning of RASnet.

On an original image with size (W, H), the coordinates for the eight landmarks are denoted as $l_1, \ldots, l_8$, respectively. A region $r \in \{1, \ldots, N\}$ is denoted by (x, y, w, h), where (x, y) is the center coordinate and (w, h) is the size. We empirically define the center coordinate and the size of each region, as shown in table 2. In particular, for global views, we select the region as the whole image.

The mask for the r-th region is denoted by $\hat{\mathbf{M}}_r$, which is the same size with the input conv features of the MA block. $\hat{\mathbf{M}}_r$ is computed by the following formula:

$$\hat{M}_r(i,j) = \begin{cases} 1, & if\ \lambda(x_r - 0.5w_r) \le i \le \lambda(x_r + 0.5w_r) \\ & and\ \lambda(y_r - 0.5h_r) \le j \le \lambda(y_r + 0.5h_r) \\ 0, & otherwise \end{cases} \quad (4)$$

where λ is the scale factor, which is used to adjust the size of the mask to fit the input feature size of MA block. Because we embed this block between the fourth and fifth conv blocks of the backbone, the value of λ is 1/16. We construct the masks using formula (4) as supervision to train the RASnet.

B. EXPRESSION RECOGNITION SUB-NET

The ERSnet has a backbone that performs feature learning and classification. The MA block fuses diversified attention to guide the discriminative expression feature learning. In particular, the MA block first learns region-specific salient features by combining region attention and channel attention and then fuses the learned features to obtain comprehensive discriminative representations. The benefits of our attention mechanism are three-fold. First, the use of region attention filters out irrelevant facial parts as much as possible, and this greatly improves the robustness of our model to against complex background variations. Second, the use of channel attention improves the model's ability to learn region-specific feature saliencies, and this aids in learning discriminative features. Third, fusing diversified attention enables the model to adapt to different expressions and individuals.

1) HYBRID ATTENTION

The hybrid attention branch contains N sub-branches, and each sub-branch specializes attention for one region. In each sub-branch, the first layer filters irrelevant regions. The next is a convolution layer that serves as a buffer, and it learns diversified features for the current region from the shared representations outputted by the fourth conv block. Region-specific channel attention is implemented with a SE block. The SE block includes squeeze and excitation operations. The former squeezes global spatial information into a channel descriptor using global average pooling. The latter captures channel dependencies using the fc layer, ReLU and sigmoid.

$\mathbf{V}$ denotes the outputs of the fourth conv block of ERSnet. The r-th sub-branch has inputs $\mathbf{V}$ and $\mathbf{M}_r$. The outputs $\mathbf{F}$ for the r-th sub-branch can be formulated as

$$\begin{aligned} \mathbf{F} &= f(\mathbf{V}, \mathbf{M}_r : \Omega) = f_{scale}(f_{SE}(\mathbf{t} : \omega), \mathbf{t}), \\ \mathbf{t} &= \omega_2(\mathbf{V} \otimes \mathbf{M}_r) + b_2 \end{aligned} \quad (5)$$

where $\mathbf{V} \otimes \mathbf{M}_r$ denotes that each channel of $\mathbf{V}$ multiplies $\mathbf{M}_r$. ω_2 and b_2 are the convolutional filter weight and bias of the buffer, respectively. $\mathbf{t}$ represents the outputs of these two layers. $f_{SE}(\cdot)$ denotes the squeeze and excitation operations with parameter ω, and it outputs activation map that is the same size with $\mathbf{t}$. The final outputs $\mathbf{F}$ are obtained by rescaling $\mathbf{t}$ with the activation map using a scale operation f_{scale}.

2) MULTIPLE ATTENTION FUSION

The salient features vary because of the different granularities considered. For example, global attention can learn global saliency, while local attention can learn fine saliency. Therefore, fusing multiple attention helps the model effectively learn discriminative features. However, because different saliencies contribute differently to expression recognition, prematurely fusing them may impact performance. Hence, the weight learning branch is built to learn the importance of different regions before the fusion occurs.

The weight learning branch inputs the features from the fourth conv block, and outputs a N-dimensional vector $\mathbf{q} = [q_1, \ldots, q_r, \ldots, q_N]$. Therefore, the outputs of MA block can be computed by the weighted sum of all the sub-branches:

$$\mathbf{F} = \sum_{r=1}^{N} q_r \mathbf{f}_r \tag{6}$$

In this way, our network can adaptively learn comprehensive discriminative features.

3) LOSS FUNCTION

We regard the FER task as a typical classification problem, and denote the ground-truth expression label for the k-th sample as a one-hot vector $\mathbf{y}_k = [y_k^1, y_k^2, \ldots, y_k^c, .., y_k^C]$, where C is the class number. The ERSnet is trained via cross-entropy loss, which is defined as follows:

$$L_E = -\frac{1}{K} \sum_{k=1}^{K} \sum_{c=1}^{C} y_k^c log\, p_k^c \tag{7}$$

where p_k^c denotes the predicted probability of the c-th category for the k-th sample.

C. TRAINING

It is empirically difficult to train the two sub-nets simultaneously because the ERSnet needs to encode high-level semantics and the RASnet needs to only encode simple spatial information. Hence, we use a two-stage training strategy. In the first stage, the RASnet is trained end-to-end and ERSnet does not update its parameters during training. In the second stage, the ERSnet is trained while RASnet does not update its parameters and serves as a mask generator for MA block.

IV. EXPERIMENT RESULTS

We evaluate the proposed method on two publicly available databases: RAF [14] and CK+ [26]. After obtaining our results, we compared them with those of state-of-the-art methods.

A. DATABASE

RAF is a widely used expression database that contains real-world facial images with large variations in illumination, occlusion, and background. Its 29,672 images are divided into single-label and two-tab subsets. The first subset was used in our experiment, and it contains 15,339 images (12,271 samples for training and 3,068 for testing). Each image is labeled with one of seven emotional categories: anger, disgust, fear, happiness, sadness, surprise and neutral.

CK+ database is collected in lab environment, and this database was also used in our experiment. It contains 327 video sequences from 123 subjects. Each sequence is labeled with one of seven expressions: 6 basic expressions (anger, disgust, fear, happiness, sadness, surprise) and contempt. In addition, each sequence displays the expression from the neutral to the peak frame. We selected only the last three frames of each sequence, resulting in 981 images. For fair comparison with the existing state-of-the-art methods [6], [15]–[17], [32], we adopted person-independent 10-fold cross-validation protocol for the CK+ database. Specifically, we arranged the IDs in ascending order and divided the database into ten subsets using sampling interval 10. One subset was used for testing, and the rest for training. The average recognition accuracy measures performance.

Both the CK+ and RAF databases have provided landmarks. In the experiments, we used affine transformation to align the images according to the landmarks, and we then cropped and resized the images to 224 × 224 resolution. The ground-truth masks for training RASnet were generated using eight landmarks (as shown in figure 4) provided by the databases.

B. IMPLEMENTATION

Our method was implemented using Caffe on GTX 2080 Ti. The dropout ratio, momentum, weight decay, and batch size were set to 0.5, 0.9, 0.0005, and 42, respectively. We began with a learning rate of 0.001. For the RAF database, the learning rate was decreased by multiplying it with 0.1 after each 5K iterations, and the total training number was set to 20K. For the CK+ databases, the learning rate was decreased by multiplying it with 0.1 after each 0.5K iterations, and the total training number was set to 1K. On both RAF and CK+ databases, a popular pre-trained model VGG-Face [27], trained for the related task of facial recognition, was used to initialize the five conv blocks of the backbone of ERSnet. Xavier was used to initialize the remaining layers.

C. PERFORMANCE COMPARISON ON THE RAF DATABASE

Table 3 shows the performance comparison between our method and state-of-the-art methods. We used VGG as backbone, and we fine-tuned it on RAF database and reported the result in table 3 as a baseline. In addition, we implemented a single attention model, referred to as VGG+SE, for comparison. This model was built by embedding a SE block between the fourth and fifth conv blocks of VGG. These two baseline methods were initialized with VGG-Face. For comparing with the methods without using transfer, we also reported the results of training from scratch. In general, using appropriate transfer can greatly improve performance.

VGG+SE obtained an accuracy of 83.28%, indicating a slight performance degradation compared to VGG fine-tuning. One possible reason is that the SE block in VGG+SE model learns attention from the whole image, thus making it

prone to dispersing attention to irrelevant facial components, as demonstrated in figure 5.

Li *et al.* [14] released the RAF database and proposed a deep locality-preserving CNN (DLP-CNN) for FER tasks. To improve the discriminative ability of the deep features, locality-preserving loss was developed by minimizing intra-class local scatters. The authors directly extracted features from the DLP-CNN to train a SVM classifier, and they obtained an accuracy of 74.20%.

Li *et al.* [10] used landmark information to extract small patches of interest and embedded PG-Units into their CNN, aiming to improve occlusion perception by learning the weights for these patches. They adopted VGG16 as a base and used the pre-trained model based on ImageNet dataset for initialization. However, their network only learnt region-level attention, and they obtained an accuracy of 83.27% on this database.

Fan *et al.* [20] proposed a multi-region ensemble CNN (MRE-CNN) framework. They cropped different local facial regions and then used paired images (i.e., the cropped region and the whole image) as the input to fine-tune VGG-Faces to obtain multiple classifiers. In the inference stage, they made an ensemble of these classifiers, resulting in an accuracy of 76.73%.

FSN [21] embedded a feature selection mechanism into an existing advanced deep architecture, i.e., AlexNet. The mechanism used pre-defined masks to locate facial regions and isolate relevant features. They used the pre-trained result on the ImageNet dataset to initialize the five convolution layers and then fine-tuned the whole network. This method could not automatically learn masks, and finally obtained an accuracy of 72.46%.

Ghosh *et al.* [2] fused individual expression, facial visual attribute, and scene information in a deep network framework for automatic group-level affect analysis. They trained a capsule network on RAF database and obtained an accuracy of 77.48% .

Our method obtained an accuracy of 85.82%, which is a significant performance gain compared to the baselines VGG fine-tuning and VGG+SE. Moreover, our method outperformed the state-of-the-art methods [10], [14], [20], [21].

Table 4 shows the confusion matrix of our method. Happiness was easily distinguishable, with a high recognition accuracy of 94.60%. Disgust and fear were the most difficult to recognize, and they both had relatively low recognition accuracies. Disgust was often misclassified as neutral, and fear was often misclassified as surprise. This may be caused by the fact that these expressions share many similar appearance features.

1) VISUALIZATION

To get a better understanding of the proposed method, figure 5 compares the visualization results (i.e., grad-cam maps and guided grad-cam maps [28]) between the proposed method and the two baseline methods. VGG fine-tuning and VGG+SE obtained messy attention as the class-specific activations dispersed to the background and other expression-irrelevant facial components. In contrast, our model learned from more explanatory regions that improved its FER accuracy, such as nose, mouth and eyes, as demonstrated in the grad-cam maps. Besides, our model learned more accurate features than the two baseline models, as demonstrated in the guided grad-cam maps.

The class activations of multiple individuals for our method are shown in figure 6, where each row shows the same expression. As can be seen, our method can learn discriminative features from expression-related regions that vary with individuals.

TABLE 3. Performance comparison on the RAF database.

Method	Accuracy(%)
Ours	85.69 / 75.59*
VGG+SE	83.28 / 73.27*
VGG fine-tuning	83.89 / 73.63*
PG-CNN [10]	83.27
DLP-CNN [14]	74.20*
MRE-CNN [20]	76.73
FSN [21]	72.46
Ghosh et al. [2]	77.48 **

* denotes training from scratch.
** denotes training from scratch or pre-training is not mentioned.

TABLE 4. Confusion matrix for the RAF database (%).

		Predicted label						
		SU	FE	DI	HA	SA	AN	NE
True label	SU	85.11	2.13	1.52	2.74	2.43	1.52	4.56
	FE	16.22	58.11	0.00	5.41	9.46	5.41	5.41
	DI	1.25	1.88	60.62	9.38	8.13	5.63	13.13
	HA	0.51	0.00	0.68	94.60	1.35	0.17	2.70
	SA	0.63	0.84	2.72	3.56	82.85	1.05	8.37
	AN	2.47	2.47	4.94	7.41	1.23	75.93	5.56
	NE	1.91	0.15	2.21	4.85	6.91	0.29	83.68

D. PERFORMANCE COMPARISON ON THE CK+ DATABASE

Table 5 reports our experimental results and shows the comparisons with the state-of-the-art methods on CK+ database. VGG+SE outperformed VGG fine-tuning. This is because the CK+ database is collected in lab environment and has few background variations. Our method achieved the high accuracy of 96.28%, and this is because fusing diversified attention enabled the model to learn more discriminative features.

Notably, the proposed method outperformed other methods, including deep learning and traditional ones [6], [11], [15], [16] that adopted the same person-independent 10-fold cross-validation protocol for the CK+ database. Liu et al. [11] proposed STM-ExpLet to learn mid-level expression representation, and obtained accuracy of 94.19%. AUDN [15] used a feature selection method to detect AU-aware receptive fields, and this mechanism was embedded into the CNNs to learn high-level features. But this method is prone to receiving expression-irrelevant fields because the true AUs are not always detected. Cai *et al.* [16] proposed using island

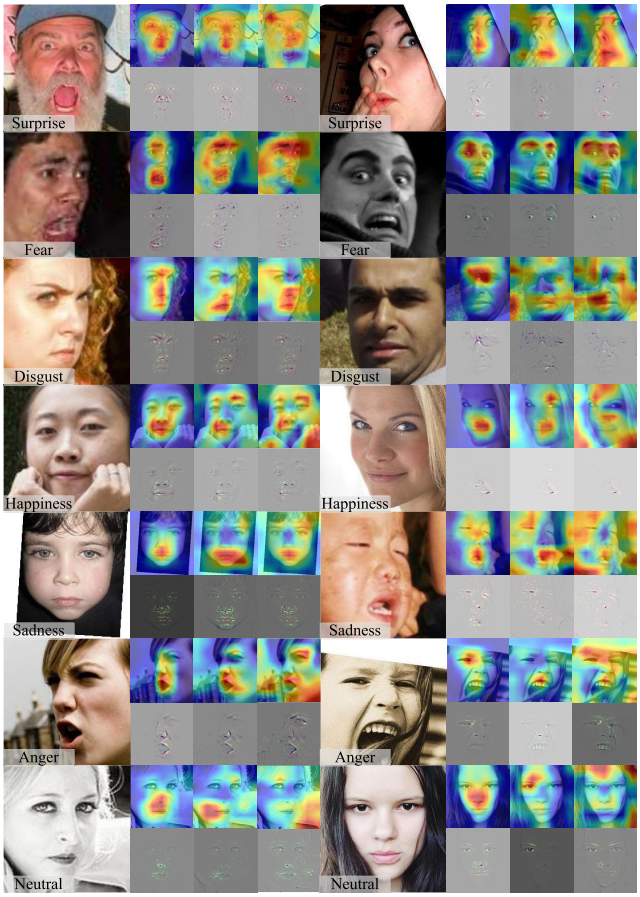

FIGURE 5. Visualization comparison for the three methods on RAF database. From left to right, the guided grad-cam maps correspond to the proposed method, VGG fine-tuning, and VGG+SE, respectively. From top to bottom, the maps correspond to surprise, fear, disgust, happiness, sadness, anger and neutral, respectively.

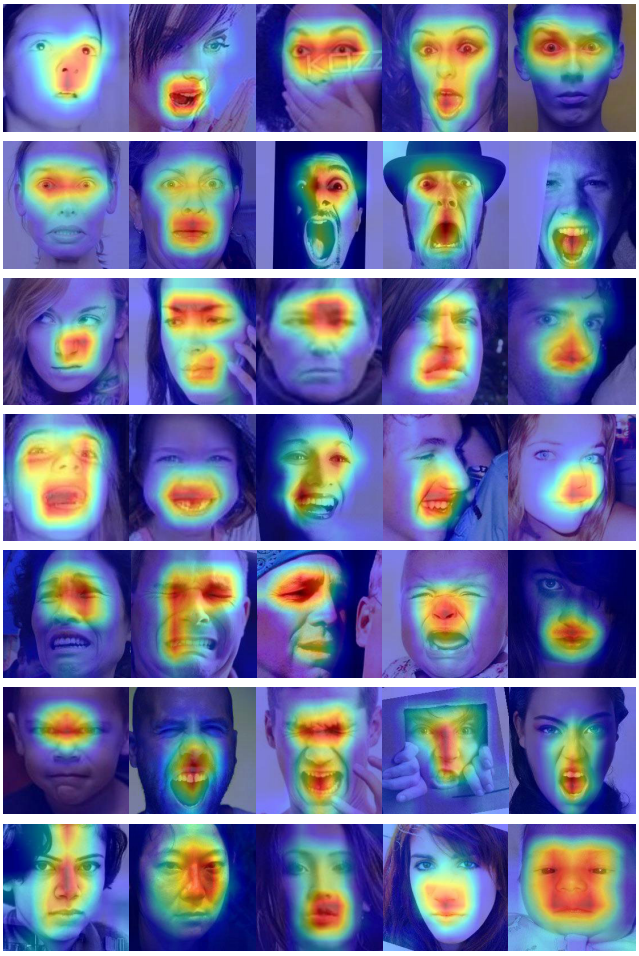

FIGURE 6. Class activations of multiple individuals on RAF database. From top to bottom, the expressions correspond to surprise, fear, disgust, happiness, sadness, anger and neutral, respectively.

loss to train CNNs, which ideally reduced the intra-class similarities and increased the inter-class similarities of the deep features, and they obtained an accuracy of 94.35%. Happy and Routray [6] extracted features from salient facial patches for expression recognition. To explore active facial patches, they detected facial landmarks via a coarse-to-fine detection method. But their method is still easy to introduce noise, thus leading to relatively poor results. Ouellet [33] implemented real-time emotion recognition using convolutional network features. Zeng *et al.* [34] compressed features by principal component analysis, and then trained an established deep sparse autoencoders (DSAE), obtaining the accuracy of 95.79%. Meng *et al.* [31] proposed an identity-aware convolutional neural network (IACNN), where expression-sensitive contrastive loss and identity-sensitive contrastive loss were built to achieve identity-invariant expression recognition. Zhao *et al.* [30] proposed a Peak-Piloted Deep Network (PPDN). PPDN used the deep features of a peak expression to supervise the intermediate feature learning of a non-peak expression. PPDN obtained accuracy of 97.30% when recognizing 6 basic expressions. Liu *et al.* [17] proposed (N+M)-tuplet clusters loss, and developed an identity-aware hard-negative mining scheme and an online positive mining scheme to learn identity-invariant features for expression recognition. Yang et al. [32] proposed De-expression Residue Learning (DeRL), which extracted the expressive component generated by the generator of GAN for recognition. Both of these two methods [17], [32] used data augmentation to augment database and obtained high accuracy. Other methods [8], [10] also adopted attention mechanism in their deep learning networks, and they obtained accuracies of 97.23% and 97.03%, respectively. However, to test a new sample, these two methods require extra facial landmark information. The method in [8] first needs to detect landmarks to generate visual fixation sequences as inputs. PG-CNN [10] needs to extract small patches using landmarks in the middle layer of the network for attention. In contrast, without using landmarks, our method can automatically predict masks for different regions, which then support the multiple attention for the expression recognition.

TABLE 5. Performance comparison on the CK+ database.

Method	Data selection	Class	Accuracy (%)
Ours	the last three frames	7†	96.28
VGG+SE	the last three frames	7†	93.11
VGG fine-tuning	the last three frames	7†	92.27
AUDN [15]	the first and the last three frames	7‡	93.70
Island loss [16]	the last three frames	7†	94.35
Happy et al. [6]	the last frame	6	94.14
STM-ExpLet [11]	sequence	7†	94.19
Jang et al. [8]	the last frame	7†	97.23
PG-CNN [10]	the first and last frames	7‡	97.03
Ouellet et al. [33]	the last frame	7†	94.40
DSAE [34]	the first and the last four frames	7†	95.79
IACNN [31]	the last three frames	7†	95.37
PPDN [30]	the last three frames	6	97.30
Liu et al. [17]	the last three frames	7†	97.10
DeRL [32]	the last three frames	7†	97.30

† denotes 6 basic expressions and contempt.
‡ denotes 6 basic expressions and neutral.

TABLE 6. Confusion matrix for the CK+ (%).

		Predicted label						
		AN	CO	DI	FE	HA	SA	SU
True label	AN	89.57	0	2.33	0	0	8.10	0
	CO	5.00	88.33	0	5.00	0	1.67	0
	DI	0	0	100.00	0	0	0	0
	FE	0	0	0	84.17	5.83	0	0
	HA	0	0	0	0	100.00	0	0
	SA	5.00	0	0	0	0	95.00	0
	SU	0	1.67	0	0	0	0	98.33

Table 6 shows the confusion matrix for our method. It can be observed that happiness and disgust were perfectly recognized. Surprise was also easier to distinguish, and it had accuracy above 98%. Fear was relatively hard for the network to recognize, and it was easily misclassified as happiness.

1) VISUALIZATION

Figure 7 shows the guided grad-cam maps for the proposed method and the two baseline methods. Although the images' backgrounds contain few variations, the two baseline methods still easily learned irrelevant features, such as hairstyles and facial outlines. Moreover, these two baseline methods seemed to treat different facial regions equally. And they both learnt features from the whole face for different expressions, without emphasizing the differences between expressions. In contrast, our method intensified the important regions while suppressing the less important ones, and learnt more accurate features, resulting in a better performance.

Figure 8 shows the class activations of multiple individuals and multiple expressions. Because in CK+ database, each individual has only 4∼5 expressions or less that are labeled, so we present the visualizations for part expressions. CK+ database contains posed expression from different individuals in a tightly controlled environment. But for the same expression, each individual's performance still has some differences. In figure 8, each column shows that our method

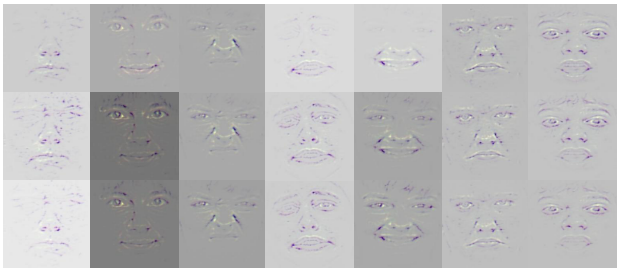

FIGURE 7. Visualization comparison for the three methods on CK+ database. From top to bottom, the guided grad-cam maps correspond to the proposed method, VGG fine-tuning, and VGG+SE, respectively. From right to left, the maps correspond to anger, contempt, disgust, fear, happiness, sadness, and surprise, respectively.

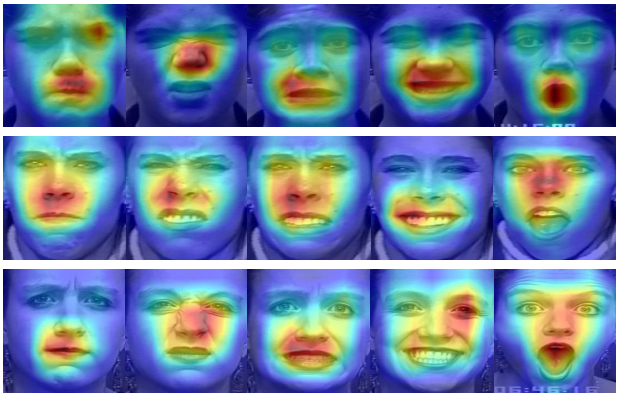

FIGURE 8. Class activations from multiple individuals and multiple expressions on CK+ database. Each row presents different expressions from one individual, and each column presents the same expression.

can learn different discriminative regions for different individuals. As can be seen from each row, different expressions have different activations. The visualization indicates that our method can learn variant attention for different expressions and individuals.

V. CONCLUSION

To improve the extraction of discriminative features for FER tasks, we have presented a novel multiple attention network that learns facial representations by simulating coarse-to-fine visual attention. The proposed network includes a RASnet and an ERSnet. The RASnet learns binary masks for automatically locating critical regions, and the ERSnet learns discriminative features by embedding a MA block that fuses multiple attention from the critical regions. The proposed method has been evaluated on two publicly available databases, RAF and CK+, and it has achieved accuracies of 85.69% and 96.28%, respectively. Comparisons with state-of-the-art methods verify that the proposed method achieves competitive or better FER performance.

REFERENCES

[1] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real time face detection and facial expression recognition: Development and applications to human computer interaction," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop*, vol. 5, Jun. 2003, p. 53.

[2] S. Ghosh, A. Dhall, and N. Sebe, "Automatic group affect analysis in images via visual attribute and feature networks," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 1967–1971.
[3] H. Gao, A. Yuce, and J.-P. Thiran, "Detecting emotional stress from facial expressions for driving safety," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 5961–5965.
[4] J. Tao and T. Tan, "Affective computing: A review," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.* Berlin, Germany: Springer, 2005, pp. 981–995.
[5] A. Yao, J. Shao, N. Ma, and Y. Chen, "Capturing AU–aware facial features and their latent relations for emotion recognition in the wild," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, 2015, pp. 451–458.
[6] S. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Trans. Affective Comput.*, vol. 6, no. 1, pp. 1–12, Jan./Mar. 2014.
[7] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, D. N. Metaxas, "Learning active facial patches for expression analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2562–2569.
[8] J. Jang, H. Cho, J. Kim, J. Lee, and S. Yang, "Facial attribute recognition by recurrent learning with visual fixation," *IEEE Trans. Cybern.*, vol. 49, no. 2, pp. 616–625, Feb. 2018.
[9] W. Sun, H. Zhao, and Z. Jin, "A visual attention based roi detection method for facial expression recognition," *Neurocomputing* vol. 296, pp. 12–22, Jun. 2018.
[10] Y. Li, J. Zeng, S. Shan, and X. Chen, "Patch–gated CNN for occlusion-aware facial expression recognition," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2209–2214.
[11] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1749–1756.
[12] L. Xu, J. Chen, and Y. Gan, "Head pose estimation using deep multitask learning," *J. Electron. Imag.*, vol. 28, no. 1, 2019, Art. no. 013029.
[13] L. Xu, J. Chen, Y. Gan, "Head pose estimation with soft labels using regularized convolutional neural network," *Neurocomputing*, vol. 337, pp. 339–353, Apr. 2019.
[14] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality–preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2852–2861.
[15] M. Liu, S. Li, S. Shan, and X. Chen, "Au-inspired deep networks for facial expression feature learning," *Neurocomputing*, vol. 159, pp. 126–136, Jul. 2015.
[16] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. Oreilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 302–309.
[17] X. Liu, B. V. K. V. Kumar, J. You, and P. Jia, "Adaptive deep metric learning for identity–aware facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 20–29.
[18] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine–tuning in deep neural networks for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2983–2991.
[19] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.
[20] Y. Fan, J. C. Lam, and V. O. Li, "Multi-region ensemble convolutional neural network for facial expression recognition," in *Proc. Int. Conf. Artif. Neural Netw.* Cham, Switzerland: Springer, 2018, pp. 84–94.
[21] S. Zhao, H. Cai, H. Liu, J. Zhang, and S. Chen, "Feature selection mechanism in CNNs for facial expression recognition.," in *Proc. Brit. Mach. Vis. Conf.*, 2018, p. 317.
[22] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.
[23] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2285–2294.
[24] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine–grained image recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5209–5217.
[25] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask–guided contrastive attention model for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1179–1188.
[26] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn–kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 94–101.
[27] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, vol. 1, 2015, p. 6.
[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 618–626.
[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[30] X. Zhao, X. Liang, and L. Liu, "Peak-piloted deep network for facial expression recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 425–442.
[31] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity–aware convolutional neural network for facial expression recognition," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 558–565.
[32] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2168–2177.
[33] S. Ouellet, "Real-time emotion recognition for gaming using deep convolutional network features," Aug. 2014, *arXiv:1408.3750*. [Online]. Available: https://arxiv.org/abs/1408.3750
[34] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, and A. M. Dobaie, "Facial expression recognition via learning deep sparse autoencoders," *Neurocomputing*, vol. 273, pp. 643–649, Jan. 2018.

YANLING GAN received the B.S. degree in computer science and technology from Guangxi Normal University, Guilin, China, in 2013, and the M.S. degree in computer applications technology from Central China Normal University, Wuhan, China, in 2017. She is currently pursuing the Ph.D. degree with the National Engineering Research Center for E-Learning, Central China Normal University. Her research interests include machine learning and pattern recognition.

JINGYING CHEN received the B.S. and M.S. degrees from the Huazhong University of Science and Technology, Wuhan, China, and Ph.D. degree from the School of Computer Engineering, Nanyang Technological University, Singapore, in 2001. She holds a postdoctoral position with INRIA, France, and a Research Fellow with the University of St. Andrews and University of Edinburgh, U.K. She is currently a Professor with the National Engineering Center for E-Learning, Central China Normal University, China. Her research interests include image processing, computer vision, pattern recognition, and multimedia applications.

ZONGKAI YANG received the B.E. and M.E. degrees from the Huazhong University of Science and Technology, Wuhan, China, in 1985 and 1988, respectively, and the Ph.D. degree from Xi'an Jiaotong University, Xi'an, China, in 1991. From 1991 to 1993, he was a Postdoctoral Researcher with the Huazhong University of Science and Technology. He is currently a Professor with the National Engineering Research Center for E-learning, Central China Normal University. His research interests include machine learning, signal processing, network communication, and information technology.

LUHUI XU received the B.S. degree in electronic and information engineering from the Qingdao University of Technology, Qingdao, China, in 2009, the M.S. degree in electronics and communication engineering from Guangxi Normal University, Guilin, China, in 2015, and the Ph.D. degree from Central China Normal University, Wuhan, China. He is currently a Lecturer with Guangxi Normal University. His research interests include machine learning and pattern recognition.