Adaptive multilayer perceptual attention network for facial expression recognition

Hanwei Liu, Huiling Cai, Qingcheng Lin, Xuefeng Li, *Member, IEEE,* Hui Xiao

***Abstract*—In complex real-world situations, problems such as illumination changes, facial occlusion, and variant poses make facial expression recognition (FER) a challenging task. To solve the robustness problem, this paper proposes an adaptive multilayer perceptual attention network (AMP-Net) that is inspired by the facial attributes and the facial perception mechanism of the human visual system. AMP-Net extracts global, local, and salient facial emotional features with different fine-grained features to learn the underlying diversity and key information of facial emotions. Different from existing methods, AMP-Net can adaptively guide the network to focus on multiple finer and distinguishable local patches with robustness to occlusion and variant poses, improving the effectiveness of learning potential facial diversity information. In addition, the proposed global perception module can learn different receptive field features in the global perception domain, and AMP-Net also supplements salient facial region features with high emotion correlation based on prior knowledge to capture key texture details and avoid important information loss. Many experiments show that AMP-Net achieves good generalizability and state-of-the-art results on several real-world datasets, including RAF-DB, AffectNet-7, AffectNet-8, SFEW 2.0, FER-2013, and FED-RO, with accuracies of 89.25%, 64.54%, 61.74%, 61.17%, 74.48%, and 71.75%, respectively. All codes and training logs are publicly available at https://github.com/liuhw01/AMP-Net.**

***Index Terms*—Facial expression recognition, Facial perceptual mechanism, Occlusion, Variant pose**

I. Introduction

EMOTIONS play an important role in human brain dynamics and have been extensively studied in the past few decades [1], [2]. As the most prominent emotionally explicit feature, facial expressions can convey information about emotions and intentions due to their adaptability and communicability [3]. With the advances of computer vision, facial expression recognition (FER) can capture the emotions of target objects and has been widely used for human-computer interaction (HCI) [4], medical diagnosis [5], and other fields.

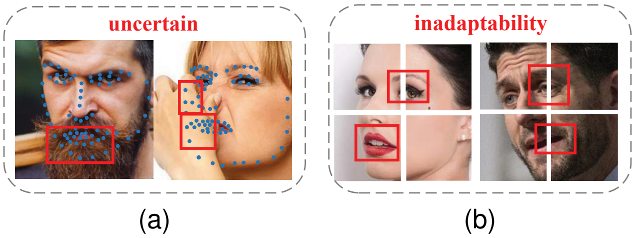


Fig. 1. Facial schematic diagram under an occlusion and variant poses. (a) Uncertain facial landmark location in landmark-based methods under occlusion, (b) Inadaptability of facial local region in image-based methods under variant poses.

Currently, FER has achieved excellent recognition results in data collection in a controlled laboratory environment, such as CK+ [6], JAFFE [7], and MMI [8]. However, the complexity and variability of real-world scenarios, such as illumination variation, face occlusion, pose variation, and other uncontrollable factors, increase recognition difficulty. Although researchers are working to increase the diversity of real-world datasets [9], [10] to improve the versatility of the model, occlusion and variant poses markedly change facial visual appearances, resulting in inaccurate feature location, imprecise face alignment or inefficient feature extraction, which make FER still a challenging. Traditional methods consider a face as a whole [11] and solve the FER problem by optimising a loss function [12], [13] or synthesizing facial expressions [14], [15] to improve generalizability. However, these methods pay less attention to the potentially diverse emotional information provided by facial details, and irregular faces caused by occlusion and variant poses also strongly affect the model's ability to extract features.

Recent studies have shown that different facial areas display diverse emotional information [16], and extracting different fine-grained features of global and local faces can mine potential key information [17] and effectively deal with information loss caused by occlusion and variant poses. Therefore, research now focuses on solving the problem of FER in the real-world situations using global and local patch methods [18], [19], [20], [21]. These studies primarily include landmark-based patches [19], [20] and image-based patches [18], [21]. Landmark-based methods can better locate facial muscle movement subregions related to emotional expression. Li *et al.* [20] proposed perceiving facial occlusion areas based on

This work was supported by the Shanghai Science and Technology Planning Project (20DZ2205900), Shanghai Municipal Commission of Science and Technology Project (19511132101), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100), and the Fundamental Research Funds for the Central Universities.

The authors are with College of Electronics and Information Engineering, Tongji University, Shanghai, 201804, China. Xuefeng Li is also with Frontiers Science Center for Intelligent Autonomous Systems, Tongji University. E-mail: liuhw1@tongji.edu.cn, caihuiling@tongji.edu.cn, 1810853@tongji.edu.cn, lixuefeng@tongji.edu.cn, xiaohui@tongji.edu.cn. Corresponding author: Prof. Xuefeng Li, Hui Xiao, Tel: +86 21-69589241, lixuefeng@tongji.edu.cn, xiaohui@tongji.edu.cn

0000–0000/00$00.00 © 2022 IEEE

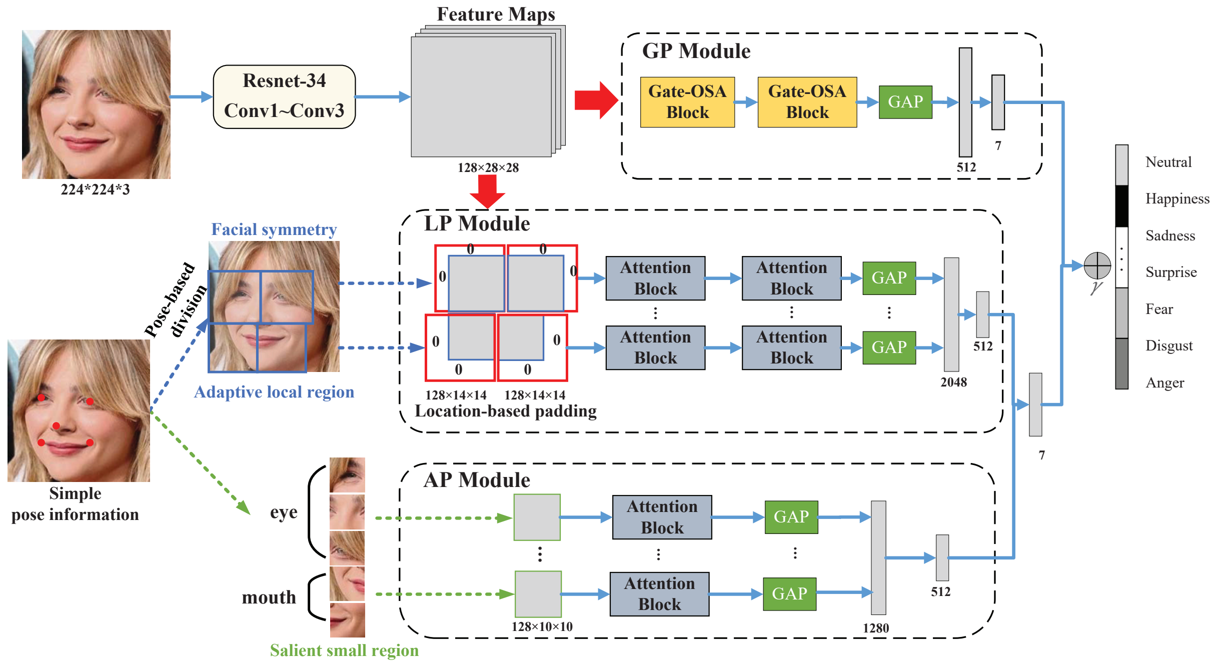


Fig. 2. Structure of the proposed method. The conv1 to conv3 layer of ResNet-34 serves as the backbone. Then, the feature maps are input into the three branch modules (GP module, LP module, and AP module) to extract different perceptual field features. The GP module is the global perception module, the LP module is the local perception module, and the AP module is the attention perception module. Finally, recognition results are obtained by fusing the feature and decision levels.

the pitches obtained from the regions of interest of 24 facial landmarks. Wang *et al.* [19] constructed a weighted mask based on 68 facial landmarks to capture global and local facial information. However, the excessive facial landmark demand relies heavily on reliable and accurate face detection and landmark tracking, and occlusion may lead to incorrect positioning of certain landmark information, as shown in Fig.1(a). Bead occlusion leads to uncertainty in landmark detection. Image-based patches can segment images into different regions at the image level to mine potential attributes. Zhao *et al.* [18] proposed a method to divide the feature map into four nonoverlapping local regions to eliminate the instability of multilandmark-based methods. Li *et al.* [21] proposed a method to focus the identifiable area using window sliding. However, the image-based method lacks adaptability to variant poses, and the same face region may be assigned to different patches under variant poses, as shown in Fig.1(b), which reduces the model's ability to learn details of irregular faces. These robustness problems limit the performance of FER in real-world situations.

Cognitive science and psychological research have shown that the human facial perception mechanism is a process from coarse to fine. For partial occlusion, facial symmetry makes it possible to capture similar emotional information in the corresponding occlusion area during emotional expression [22]. The eyes and mouth regions convey more emotional information due to marked local muscle changes [23]. The perceiver's visual system also focuses more on the eye and mouth areas [24], [25]. Therefore, to improve robustness and the performance of FER in real-world situations, we innovatively propose an adaptive multilayer perceptual attention network (AMP-Net) that is inspired by the facial attributes and the facial perception mechanism of the human visual system. AMP-Net extracts different fine-grained emotional features from global, local, and salient facial regions to learn the diversity and key information of facial emotions under real-world scenarios (see Fig. 2). AMP-Net has three different perception domain modules. The proposed local perception module (LP module) is robust to occlusion and variant poses, and can thus guide the network to focus on multiple finer and distinguishable local patches based on facial attributes and to learn diverse potential information through the adaptive local region method and attention blocks. The LP module achieves a reasonable distribution of patches under variant poses, and the obtained local patches also exhibit facial symmetry, which can provide similar information for occluded parts. Conversely, in the global perception module (GP module), the proposed gate one-shot aggregation (gate-OSA) block can enhance features with different receptive fields in the global perceptual field. In addition, to avoid information loss caused by the inaccurate positioning of the model for key regions, we also use an attention perception module (AP module) to supplement the key texture details of eye and mouth regions with high emotional correlation based on prior knowledge to learn the differences in facial expressions. Therefore, the robustness and effectiveness to occlusion and variant poses can be increased through different levels of perceptual fields.

The contributions of this study include the following:

- We propose an adaptive multilayer perceptual attention network (AMP-Net) based on the facial attributes and facial perception mechanism that can adaptively capture the diversity and key information from global, local, and salient facial regions to improve the robustness of FER in real-world situations.

- We designed a local perception module that is robustness to occlusion and variant poses, and can effectively extract potential information from different facial regions.
- A global perception module is designed to obtain features with different receptive fields, and an attention perception module supplements salient emotional features based on prior knowledge.
- Extensive experimental results on real-world datasets (i.e., RAF-DB, AffectNet, SFEW 2.0, FER-2013, and FED-RO) demonstrate that AMP-Net achieves state-of-the-art expression recognition performance. In particular, we also perform experiments with occlusion and variant-pose datasets to show the improved robustness of the proposed method.

II. Related Work

In this section, we mainly present related works on FER under human visual and computer vision, aiming to inspire new FER methods by understanding the human visual perception mechanism under normal, occlusion, and variant poses before reviewing computer vision-related technologies under the same situation.

A. Human vision FER

The human visual system can quickly capture others' emotions even in complex environments. Therefore, researching new recognition methods based on human visual perception mechanisms is an important way to improve FER performance under occlusion and variant poses.

For visual perception under occlusion, Halliday *et al.* [26] conducted an experiment to determine emotions in static facial images of four different occluded areas (forehead and eyebrows, nose and cheeks, eyes, and mouth), which showed that subjects could accurately identify emotions from limited information. Additionally, the mouth and eyes were the two most critical areas for identifying real emotions. Roberson *et al.* [25] compared the effects of different masks, such as simulated sunglasses on facial expressions, and found that the ability to decode facial emotions considerably decreased when the masks were occluded by the eyes and mouth. Yan *et al.* [27] found that anger, fear and sadness were easier to recognise from the upper facial area, while disgust and happiness were easier to recognise from the lower facial area by occluding upper and lower facial areas.

For visual perception under variant poses, Busin *et al.* [28] observed asymmetry deviation in human emotion recognition by observing the 45° left and right sides of the expresser by the perceiver. The right face required more fixations than the left face. In addition, related studies [29], [30] have also shown that emotional expression on the left face is more active than that on the right face.

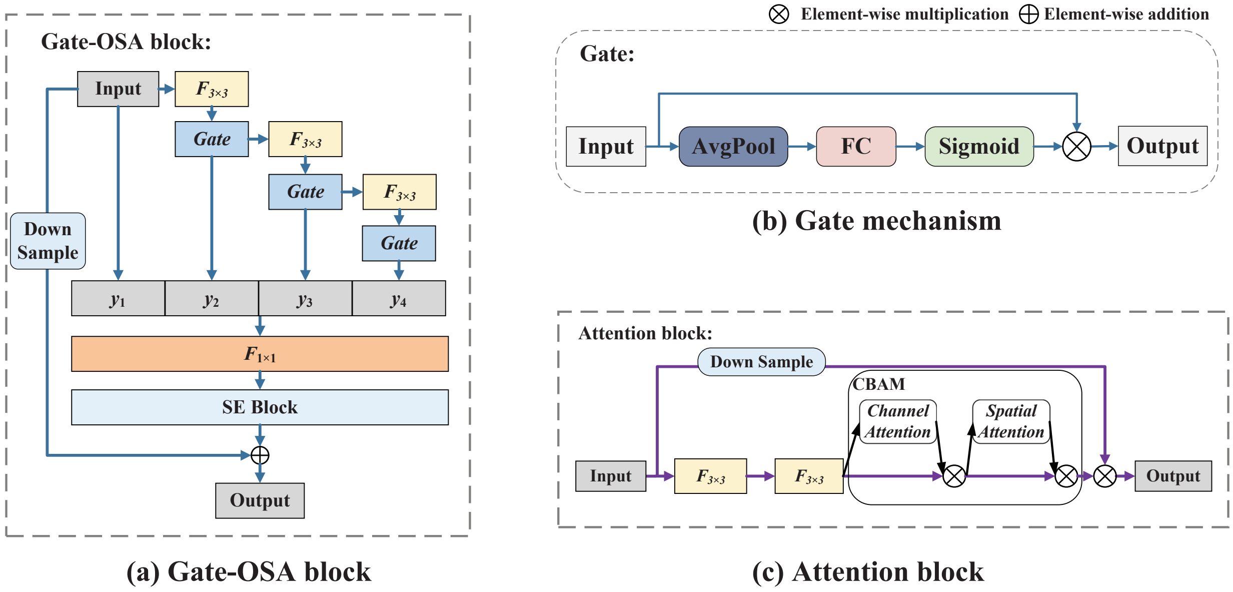
For visual perception under normal faces, Duncan *et al.* [22] quantified the visual system experiment in a bubble aspect and found that individual differences were primarily evident in the prediction of the eye area. Only the vertical information when the mouth was open was shown to be effective. Caldara *et al.* [31] studied the eye movements of observers from different cultural backgrounds in the FER task and found that although human perception and determination of facial expressions were different due to experience and environmental factors, they were more inclined to focus on eye and mouth regions. Studies [32], [33] also showed correlations of upper facial features with fear, sadness, and anger and lower facial features with surprise, disgust, happiness, and neutrality. When recognizing sadness, primarily the eyes, eyebrows, and mouth provided useful information [34]. For fear recognition, people primarily fixate on the eyes, and the mouth region can provide additional information [35]. Therefore, based on the prior knowledge of human visual FER, we propose the AMP-Net method to acquire different fine-grained facial features of global, local, and salient regions to improve the effectiveness of FER in real-world situations.

B. Computer vision FER

Occlusion and variant pose are two key FER issues in real-world scenarios. A face is likely to be occluded by sunglasses, hats, scarves, and other things, which markedly changes the facial visual appearance, and a variant poses leads to partial information loss and inaccurate positioning. Previous methods to deal with occlusion and variant poses can be categorised into two parts: holistic-based methods and pitch-based methods.

Holistic-based methods treat the face as a whole and typically solves occlusion and variant-pose issues based on feature reconstruction of geometry [36] texture [37] or improvement of loss function [12], [38] and synthesis of facial expression [14], [15]. Zhang *et al.* [36] combined the iterative closest point (ICP) algorithm and fuzzy C-means to construct a facial point detector and reconstructed 54 facial points of occluded and variant poses. Xie *et al.* [12] proposed a new triplet loss based on class-pair margins and multistage outlier suppression to enhance interclass separability and intraclass compactness of network features. Xi *et al.* [14] performed facial expression recognition and facial image synthesis simultaneously based on a generative adversarial network (GAN) to ease the overfitting problem in the FER task. However, these methods pay less attention to the potentially diverse emotional information provided by facial details, and irregular facial images caused by occlusion and variant poses also markedly affect the performance of global facial feature extraction.

Pitch-based methods extract facial subregions as regions of interest and assign different attention weights, and primarily include landmark-based [19], [20], [39], [40] and image-based [18], [21], [41] methods. Zhang *et al.* [40] proposed a Gabor-based finite element template for FER analysis based on the occlusion of the eyes, mouth, and glasses, and randomly placed blocks. In a recent study, Li *et al.* [20] perceived occlusion regions through a convolutional neural network (CNN) with an attention mechanism, gACNN focused on the global facial representation, and pACNN detected the occlusion problem in the regions of interest of 24 facial landmarks through an occlusion attention mechanism. Wang *et al.* [39] proposed a region attention network (RAN) and used fixed position cropping, random



(a) Gate-OSA block

(b) Gate mechanism

(c) Attention block

Fig. 3. Gate-OSA and Attention block are used in AMP-Net. (a) The gate-OSA block is used in the GP module, (b) The gate mechanism (gate) is used in the gate-OSA block, (c) The attention block is used in the LP module and AP module. F denotes convolutional layers, and y denotes output feature maps.

cropping, and landmark-based cropping to capture important facial pitches to solve occlusion and variant poses. Zhao *et al.* [18] proposed a visual information processing method based on a global, multiscale, and local attention network (MA-Net) in a real-world environment, and evenly divided facial images into four blocks to guide the network to focus on the local salient features. However, excessive facial landmark requirements may lead to inaccurate landmark detection, and image-based methods lack adaptability to variant poses, which limits their ability to mine facial details. Different from these methods, the proposed AMP-Net is robust to occlusion and variant poses, and requires fewer facial landmarks, which can adaptively acquire finer and distinguishable local regions under variant poses and supplement global and key region information to obtain potential features.

III. Methods

A. Overview

We propose an adaptive multilayer perceptual attention network (AMP-Net) to address occlusion and variant-pose issues. As shown in Fig. 2, AMP-Net consists of three components: the GP module, the LP module, and the AP module. The network takes a facial image as input. First, the conv1 to conv3 layer of ResNet-34 [42] serves as the feature pre-extractor to output $128 \times 28 \times 28$ feature maps. Then, the feature maps are input into the three-branch module to extract different perceptual field features, and the FER results are obtained by fusing feature-level and decision-level features.

B. GP module

The GP module aims to learn deeper global facial features in different receptive fields within the global perceptual domain. The one-shot aggregation (OSA) block [43], as a variant of DenseNet [44], aggregates all previous layers into the last layer in a relatively sparse manner and obtains rich receptive field information through feature reuse, which can effectively reduce feature redundancy caused by heavy dense connections in the DenseNet network.

To enhance available features, we design the gate-OSA block, as shown in Fig.3(a). Each 3×3 convolutional layer in the gate-OSA block is connected to a gating mechanism (gate) to learn the channel correlation, as shown in Fig.3(b). The gating mechanism compresses the output of the convolutional layer to the channel dimension through the Avgpool layer and is then transformed by a fully connected (FC) layer with a sigmoid activation function α to derive the weight of the channel. Then multiplying the output of the convolutional layer with the channel makes the channel with higher correlation have a higher weight, and the lower weight suppresses the channel with a lower correlation. The gate layer can be formulated as follows:

$$F_G(x_g) = \alpha((FC(AvgPool(x_g)))) \bigotimes x_g \qquad (1)$$

where x_g is the input of the gate layer, F_G is the output of the gate layer, and $\otimes$ is element-wise multiplication. Each gate layer in the gate-OSA module has two types of connections. One type of connection obtains feature information with a larger receptive field through alternate series connections of n 3×3 convolutional layers with gate layers. The output of each convolution and gate layer is the same $C_1 \times W \times H$ feature map, where C_1 is the number of channels in the feature map. The other type of connection connects the output of each gate layer with to last output layer to obtain the $C_2 \times W \times H$ feature map, where $C_2 = 128 + C_1 \times n$, and thus obtain feature information about different receptive fields. Then, a 1×1 convolution reduces the dimension of the $C_2 \times W \times H$ feature map to $C_3 \times W \times H$ so that the model can train a deeper network. Then, the SE block [45] is added to further enhance the features. Finally, the input of the gate-OSA block is added to the output through the down sample layer to increase the short-circuit connection, and the information loss is reduced through the feature multiplexing of multilayer receptive field information.

In the proposed network, each gate-OSA block has a gate layer serial connection of $n = 3$. The GP module has three gate-OSA blocks connected in a series, with $C_1 \in \{128, 144, 160\}$ and $C_3 \in \{256, 384, 512\}$. The GP module receives a $128 \times 28 \times 28$ feature map as input and outputs a $512 \times 7 \times 7$ feature map through three gate-OSA blocks, where each gate-OSA block halves the feature map size. Finally, a global average pooling (GAP) layer is connected to obtain a 512-dimensional global perception feature vector. GP module obtains global information with different receptive fields through multiple convolutional layer feature multiplexing and improves the performance of FER feature extraction in the global scope. A comparison of the settings is shown in ablation experiments IV-C.

C. LP module

Local facial information can provide more robust emotional feature information for occlusion and variant poses, and the selection of local patches seriously affects the model's ability to mine potential features. We propose an LP module based on facial attributes that is robust to occlusion and variant poses, and can adaptively guide the network focus on multiple finer and distinguishable local patches, improving the ability to learn potentially diverse facial emotions and eliminating the incorrect positioning that may be caused by multiple landmarks and the low adaptability caused by image-based patch methods.

Based on the following knowledge, the upper and lower facial parts convey different emotional information [23] and facial symmetry can provide similar feature information for the occlusion area [22]. Therefore, in this study, the LP module first uses pose-based division (PBD) and location-based padding (LBP) methods to divide face into four subregions with facial symmetry: upper left, upper right, lower left, and lower right. Then facial organs, such as the eyes and mouth, are allocated to the corresponding subregions to ensure the effectiveness of local region allocation under occlusion and variant poses. Local patches with facial symmetry can also provide similar emotional information for the occluded parts, reducing the impact of a partially missing face. In addition, the PBD can adaptively identify the effective facial range of different patches, eliminate the interference of the redundant parts of the image, and then mine the potentially diverse information of different facial subregions through the attention block. The details of the PBD and LBP methods are as follows:

1) Pose-based division (PBD): We first use the RetinaFace [46] facial landmark detector with occlusion robustness to extract five key points of eyes, nose, and mouth in facial maps $R \in (r, r)$ as as simple pose information, as shown in Fig. 4(a), where $P_{Eye1} = (x_{eye1}, y_{eye1})$, $P_{Eye2} = (x_{eye2}, y_{eye2})$, $P_{Nose} = (x_{nose}, y_{nose})$, $P_{Mouth1} = (x_{mouth1}, y_{mouth1})$, $P_{Mouth2} = (x_{mouth2}, y_{mouth2})$, and r represents the maximum pixel point of the image.

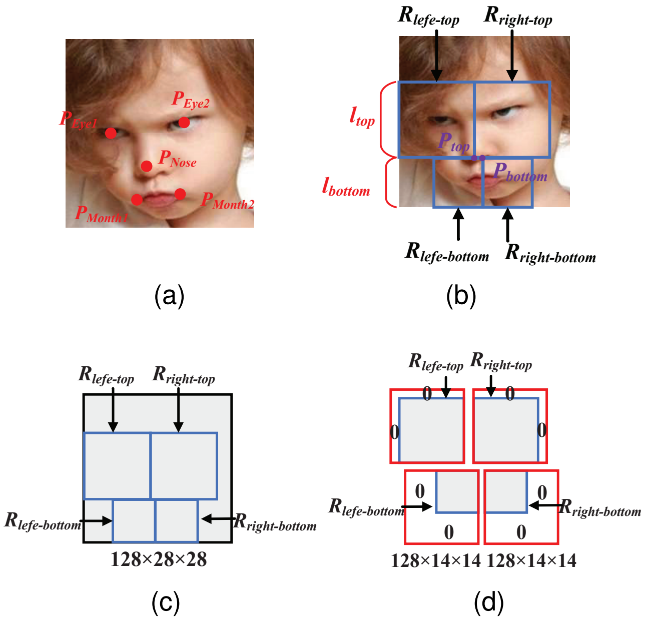


Fig. 4. The pose-based division (PBD) and location-based padding (LBP) schematic. (a) Five facial key points, (b) Four facial subregions: $R_{left-top}$, $R_{right-top}$, $R_{left-bottom}$, $R_{right-bottom}$, (c) Mapping to feature maps, (d) The LBP to obtain the uniform size of the subregion feature maps.

According to the position of the nose on the Y axis, the face is divided into top and bottom parts, and the left and right parts are divided, respectively according to the positions of the eyes and mouth on the X axis. A total of four facial subregions are obtained, meaning two division points are described by follows: $P_{top} = (x_{top}, y_{center})$ and $P_{bottom} = (x_{bottom}, y_{center})$, as shown in Fig. 4(b), where:

$$\begin{cases} y_{center} = y_{nose}, \\ x_{top} = (x_{eye1} + x_{eye2})/2 \\ x_{bottom} = (x_{mouth1} + x_{mouth2})/2 \end{cases} \tag{2}$$

In addition, to reduce the interference of useless facial information under different facial poses, we define the left and right subregions at the top and bottom based on facial symmetry as square regions of the same size with lengths l_{top} and l_{bottom}, to extract useful facial information. l_{top} and l_{bottom} are defined as the minimum distance between the division point and upper or lower image boundaries, respectively, and l_{top} and l_{bottom} can be formulated as follows:

$$\begin{cases} l_{top} = min(x_{top}, y_{center}, r - x_{top}), \\ l_{bottom} = min(x_{bottom}, r - y_{center}, r - x_{bottom}) \end{cases} \tag{3}$$

The final four facial subregions: $R_{left-top}$, $R_{right-top}$, $R_{left-bottom}$, $R_{right-bottom}$, can be obtained by the formula as follows:

$$\begin{cases} R_{left-top} \in [(P_{top} - l_{top}) : P_{top}], \\ R_{right-top} \in [P_{top} : (P_{top} + l_{top}) \\ R_{left-bottom} \in [(P_{bottom} - l_{bottom}) : P_{bottom}] \\ R_{right-bottom} \in [P_{bottom} : (P_{bottom} + l_{bottom}) \end{cases} \tag{4}$$

Using this method, a reasonable allocation of facial local regions is ensured under different poses. As shown in Fig. 4(b), the left eye is in $R_{left-top}$, the right eye is in $R_{right-top}$, the

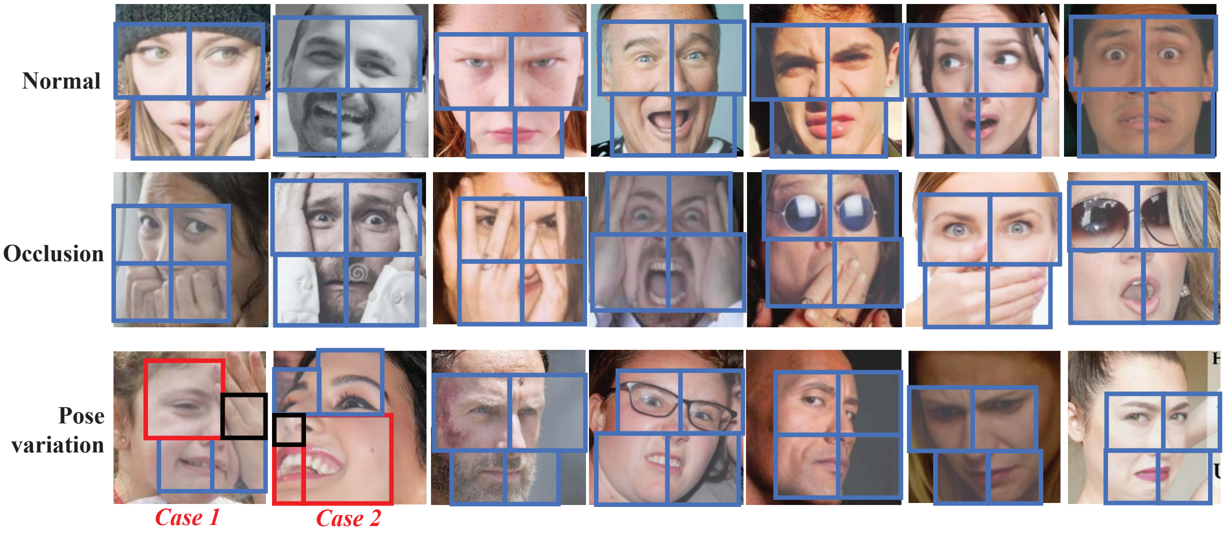


Fig. 5. Facial subregion in LP module under normal, occlusion, pose variation cases on the RAF-DB and AffectNet dataset. The blue region is the finer subregion; and the first, second, and third rows represent the facial images under normal, occlusion, and variant poses, respectively.

left lip is in $R_{left-bottom}$, and the right lip is in $R_{right-bottom}$. The following two special cases can occur:

1) When $l < r/3$, where $l = l_{top}$ or l_{bottom}, a small local region may cause the loss of some important information. Therefore, in this study, the length l_1 of a symmetric subregion with a subregion of length l is the maximum length of the division point from the image boundary in this direction of the symmetric subregion, which is less than $r/2$. As shown in *Case 1* of Fig. 5, the length of the black subregion is $l < r/3$; therefore, the length of the symmetric subregion is $l_1 = r/2$.
2) When the subregion does not contain the corresponding eyes or mouth key points, the subregion is defined as a rectangle with width l and length l_2, where l_2 is the maximum length less than $r/2$ of the division point distance from the image boundary along the vertical direction of l. The symmetrical subregion is a square with length l_2, as shown in *Case 2* of Fig.5, where the black subregion does not contain the left lip, while the modified red subregion contains the left lip.

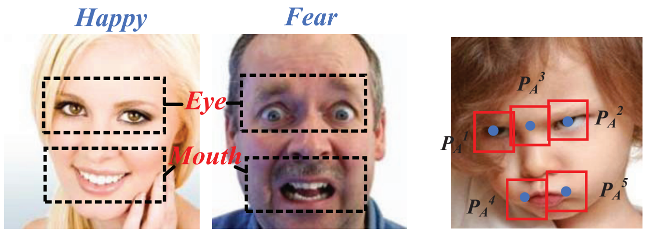
Fig. 5 shows the pose-based division method results under normal, occlusion, and pose-variation conditions. The proposed method adaptively allocates facial organs, such as the eyes and mouth, to their corresponding subregions under different facial conditions, ensuring the rationality of the facial local region distribution and eliminating the influence of invalid regions, such as hair, on feature extraction. Finally, the four facial subregions are mapped to the $128 \times 28 \times 28$ feature map, as shown in Fig.4(c).

2) Location-based padding (LBP): The output of the pose-based division method is the local region feature maps with different sizes, as shown in Fig.4(c). Because the input of the convolutional layer should be the feature map of the same size, a location-based padding method is proposed to retain the integrity of the feature information and the effective input of the convolutional layer, as shown in Fig.4(d). The four local region feature maps in the left-top, right-top, left-bottom, and right-bottom directions are padded to a fixed size of $128 \times 14 \times 14$ by value=0 in the left-top, right-top, left-bottom, and right-bottom directions. The LBP method effectively unifies all feature map sizes without changing the data structure of the feature maps, retains the position information of subregions in the corresponding face direction with fill value=0, and increases the robustness of local feature extraction under facial variant poses. The advances of the local patch acquisition method are demonstrated in ablation experiments IV-C.

Based on these steps, four local region feature maps with $128 \times 14 \times 14$ are obtained and input into the attention block in parallel. The attention block is shown in Fig.3(c), which contains two 3×3 convolution layers and a lightweight convolutional block attention module (CBAM) [47] to weigh both channel and space dimensions, making the model pay more attention to emotion-related regions and feature channels. The channel attention M_c and spatial attention M_s modules in the CBAM are set as sequential serial connections. Finally, the input feature map is added to the output feature map of CBAM and enhances the features through feature multiplexing based on short-circuit connections. The attention block can be formulated as follows:

$$F_A(x) = M_s(M_c(f(x))) \bigotimes x \tag{5}$$

where x is the input of the attention block, f represents two 3×3 convolution layers, and F_A is the output. In the LP module, two parallel attention blocks are used to extract attention features and output four $128 \times 7 \times 7$ feature maps, where the first attention block reduces the 14×14 feature map to 7×7 through down sample and the first 3×3 convolution layers. Then, the GAP layer is connected to obtain 4×512 feature vectors and spliced into 2048-dimensional feature vectors. After dimensionality reduction of the FC layer to 512 dimensions, the final facial local perception emotion feature



(a) Facial expression (b) Specific regions

Fig. 6. Specific areas of eyes and mouth. (a) Eyes and mouth show a substantial emotional correlation in emotion expression, (b) Five subregions related to the eyes and mouth as specific regions in the AP module.

is obtained. The LP module extracts finer subregion features based on simple pose information, reduces the interference of invalid regions to the model, and improves FER performance for occlusion and pose variation by learning the potential diversity facial emotions.

D. AP module

Based on the prior knowledge that the eyes and mouth show a substantial emotional correlation in emotion expression and recognition perception, as shown in Fig.6(a), the AP module is designed to extract the key texture details of salient areas of the eyes and mouth through the attention network, paying more attention to small-scale areas with important emotional features. This result is used as supplementary information to eliminate feature lack that may be caused by ignoring certain important areas.

The AP module obtains five subregions related to the eyes and mouth based on facial key points. As shown in Fig.6(b), the centre points P_A of each region are as follows:

$$\begin{cases} P_A^1 = P_{Eye1}, \\ P_A^2 = P_{Eye2} \\ P_A^3 = (P_{Eye1} + P_{Eye2})/2 \\ P_A^4 = P_{Mouth1} \\ P_A^5 = P_{Mouth2} \end{cases} \tag{6}$$

where P_A^1, P_A^2, and P_A^3 are the centres of the left eye, right eye, and eyebrow, respectively; and P_A^4 and P_A^5 are the centres of the left lip and right lip, respectively. Then the centre points of the five subregions are mapped to a $128 \times 28 \times 28$ feature map to obtain five feature map regions with $128 \times L \times L$.

The AP module inputs five feature maps into the parallel attention block to obtain $128 \times L/2 \times L/2$ feature maps with different channel and regional attention weights. Then the GAP layers are connected to obtain a 5×256-dimensional feature vector and spliced into 1280 dimensions, which is finally reduced to 512 dimensions by the FC layer as the final attention perception emotion feature. The AP module can capture the features of eyes, eyebrows, and mouth regions with substantial emotional information used as supplementary information for the network to ensure robust feature extraction. In the experiments of this study, we set $L = 10$ as the optimal value.

E. Fusion strategy

In this study, the LP module and AP module are fused at the feature-level to guide the model to pay more attention to salient regions without occlusion. The fusion of the GP module and feature-level fusion results at the decision-level can obtain features in different perception domains to ensure that the model performs robustly in facial-occlusion and pose-variation conditions.

For specific implementation, the 512-dimensional features provided by the LP module and AP module are spliced into 1024-dimensional emotional features, and the FC layer is connected to output c-dimensional vectors, which can be formulated as $z_{LA} = \{z_1, z_2, ..., z_C\}$ to achieve feature-level fusion, where c is the number of emotional categories. Then, we connect the z_{LA} and c-dimensional vector z_G output by the GP module under the FC layer to train the model through the loss function L as follows:

$$L = \lambda L_{GP} + (1 - \lambda) L_{L\_AP} \tag{7}$$

where L_{GP} is the output loss of the GP module, L_{L_AP} is the output loss of the feature fusion result, and λ is a hyperparameter used to balance L_{GP} and L_{L_AP}. In the experiment, L_{GP} and L_{L_AP} are calculated by minimising the cross entropy loss, which can be formulated as follows:

$$L_M = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{c} y_j log \hat{y_j} \tag{8}$$

where N is the number of samples, $\hat{y_j}$ is the predicted result, y_j is the true result, and $M \in (GP, L\_AP)$ are the two inputs of the decision-level.

IV. EXPERIMENTS

A. Datasets

To make fair comparisons with previous studies, we perform experiments on five popular real-world facial expression datasets (RAF-DB [9], AffectNet [10], SFEW 2.0 [48], FER-2013 [49] and FED-RO [20]), and occlusion and variant-pose test sets [39] (Occlusion-RAF-DB, Occlusion-AffectNet, Occlusion-FERPlus, Pose-RAF-DB, Pose-AffectNet, and Pose-FERPlus).

RAF-DB is a large-scale facial dataset with 30,000 facial images and is marked by approximately 40 annotators. In these experiments, we only use images with 6 basic emotions plus neutral emotions, includes 12,271 training images and 3,068 test images.

AffectNet is currently the largest facial emotion dataset, with more than 40 W images that are manually annotated into 11 discrete emotion categories can contain valence and arousal dimension emotions. In these experiments, images with 7 and 8 classes of emotions are used for testing, and data balance processing is used to address the quantitative differences between different emotion categories in the training samples. The experiments use the 7 basic emotions plus neutral, which includes 70,181 training images and 3500 test images; the 8 types of emotions are added with contempt and include 73,931 training images, and 4,000 test images.

SFEW 2.0 is created by selecting static frames from the AFEW database based on key frames and contains 7 types of emotion labels, 958 training images, 436 verification images, and 372 test images. In these experiments, because the emotion label with the test set could not be obtained, the verification set is used to evaluate the FER.

FER-2013 contains approximately 30,000 greyscale images with a size of 40×40 and seven emotion categories. We select 28,709 images as a training set and 3,589 for the public test set images to evaluate recognition performance.

FED-RO is a set of facial occlusion datasets that is searched through by the Bing and Google search engines. Images that are duplicated with the RAF-DB and AffectNet datasets are removed to obtain a total of 400 face images with seven emotion categories.

Occlusion-RAF-DB, **Occlusion-AffectNet**, and **Occlusion-FERPlus** contain images with facial occlusion collected in the validation set of AffectNet, the test set of RAF-DB, and the test set of FERPlus, respectively, which include 683, 735, and 605 images, respectively.

Pose-RAF-DB, **Pose-AffectNet**, and **Pose-FERPlus** contain images with facial-pose changes that are in the validation set of AffectNet, the test set of RAF-DB, and the test set of FERPlus, with 1949, 1248, and 1171 images that contain facial poses greater than 30°, and 1,171, and 958, 558, and 634 images with facial poses greater than 45°, respectively.

B. Implementation details

For all facial images, Retinaface [46] is used to extract five facial key points of the eyes, nose, and mouth and intercepts the facial area with a pixel size of 224×224. Random flip and translation as well as random changes in brightness, contrast, and saturation for data enhancement. To make fair comparisons with previous studies, we use ResNet-34 [42] as the backbone of the proposed method. For the RAF-DB, AffectNet, and FER-2013 datasets, first pretrained on the large-scale face recognition dataset VGGFace2 [50] and then fine-tuned. The SFEW 2.0 dataset was pretrained on the RAF-DB dataset and then fine-tuned; FED-RO, Occlusion-AffectNet, Occlusion-RAF-DB, Occlusion-FERPlus, Pose-AffectNet, Pose-RAF-DB, and Pose-FERPlus use the same settings as RAN [39]. By default, the region size L is set as 10, and the hyperparameter is set as $\lambda = 0.5$. The proposed method is implemented on the GeForce RTX 3090 Ti platform using the PyTorch toolbox [51]. The minibatch size is set to 350 with a momentum of 0.9 and a weight decay of 0.0001. The learning rate starts at 0.1 and decreases by 10 after 20 epochs. We train the model for a total of 100 epochs. Random gradient descent (SGD) was used as the optimisation algorithm. The number of parameters of AMP-Net is 105.67 M, and the number of FLOPs is 1.69.

C. Ablation experiments

To verify the effectiveness of AMP-Net, ablation experiments are performed on the GP module, the LP module, the AP module, region size L, and hyperparameter λ. The representative real-world datasets include RAF-DB, AffectNet-7 (7 classes) and FED-RO datasets, which are verified without pretraining.

TABLE I
EVALUATION ON A SINGLE MODULE ON RAF-DB, AFFECTNET, AND FER-RO **WITHOUT PRE-TRAINING**, AND GP DENOTES GP MODULE, LP DENOTES LP MODULE, AND AP DENOTES AP MODULE

Modules	RAF-DB	AffectNet-7	FED-RO
ResNet-32	84.03	60.82	61.50
GP (OSA)	86.08	61.28	64.75
GP (Gate-OSA)	86.32	61.65	65.25
LP	86.02	62.50	64.75
AP	84.33	61.32	62.00

TABLE II
EVALUATION OF EACH MODULE IN OUR AMP-NET ON RAF-DB, AFFECTNET, AND FER-RO **WITHOUT PRE-TRAINING**

Dataset	GP	LP	AP	Acc. (%)
RAF-DB	✓	✓		87.23
	✓		✓	86.77
		✓	✓	86.67
	✓	✓	✓	**88.06**
AffectNet-7	✓	✓		62.79
	✓		✓	62.33
		✓	✓	62.80
	✓	✓	✓	**63.23**
FED-RO	✓	✓		67.50
	✓		✓	66.75
		✓	✓	67.00
	✓	✓	✓	**68.25**

1) GP module: We analyse the FER performance of the GP module and the gate-OSA block, and the recognition results of each module are shown in Table I. The FER results of the GP module using the Gate-OSA are 2.29%, 0.83%, and 3.75% higher than ResNet-32 on the RAF-DB, AffectNet-7, and FED-RO datasets, respectively. These results show that the GP module can learn the global facial features through multilayer feature reuse more effectively. To verify the performance of the gate-OSA block, the gate-OSA block in the GP module is modified to become an external OSA block. Results show that the gate-OSA block improved by 0.24%, 0.37%, and 0.5% compared with OSA on the RAF-DB, AffectNet-7, and FED-RO datasets. Experimental results show that the gate-OSA block can learn the emotional characteristics of important weight channels more effectively to improve the FER performance.

2) LP module: We then evaluate the performance of the LP module. In the single module, FER results are shown in Table I, and LP module achieves excellent FER performance by extracting finer facial subregion features, which are 1.99%, 1.68%, and 3.25% higher than ResNet-32 in the RAF-DB, AffectNet-7, and FDE-RO datasets. The LP module also achieves the highest FER result (62.5%) in the recognition of AffectNet-7. In addition, the decision-level fusion of the LP module and the GP module considerably improves recognition performance by 1.21%, 0.29%, and 2.75% and 0.91%, 1.14%, and 2.25% on RAF-DB, AffectNet-7, and FED-RO, respectively, compared with the two single modules, as shown in Table II.

To explore the impact of the local feature selection strategy on the LP module, we design four different schemes for

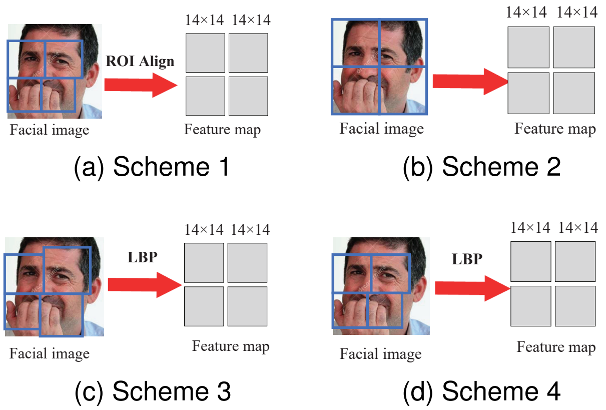


Fig. 7. Schematic diagram of different local feature selection schemes on the LP module: (a) PBD and ROI Align to pool into a uniform size; (b) Image equally divided into four parts; (c) Modified PBD and LBP methods; and (d) Proposed method.

TABLE III
EVALUATION OF DIFFERENT LP MODULE SCHEMES ON RAF-DB WITHOUT PRE-TRAINING

Schemes	Acc. (%)
Scheme 1	79.35
Scheme 2	84.65
Scheme 3	85.27
Scheme 4	**86.02**

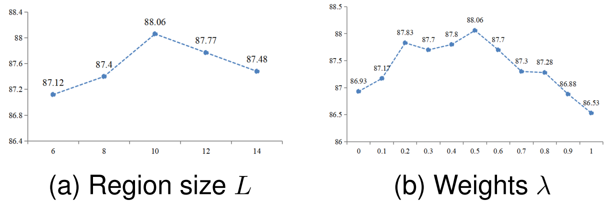


Fig. 8. The impacts of the region size L and the number of weights λ on the performance of the RAF-DB.

FER performance comparison, as shown in Fig.7. For scheme 1, pose-based division is first performed to obtain the local region, and ROI Align [52] is used to pool local regions with different sizes into four $128 \times 14 \times 14$ feature maps with uniform size, which are used as the input of the attention blocks (see Fig.7(a)). ROI Align uses bilinear interpolation to convert the feature aggregation process into a continuous operation and can pool feature maps with different sizes into the same size. For scheme 2, we divide the output feature map of the Backbone into four nonrepetitive feature maps with $128 \times 14 \times 14$ as the input of the attention blocks, which is the same setting as MA-Net [18], as shown in Fig.7(b). For scheme 3, we modify the strategy of pose-based division by changing the subregion length l to the shortest distance between the image boundary in the corresponding subregion direction and the division point. If l is greater than $r/2$, then we let $l = r/2$. Then four feature maps with $128 \times 14 \times 14$ are obtained through mapping and location-based padding as the input of the attention blocks, as shown in Fig. 7(c). For scheme 4, the proposed method is used to obtain regional features as the input of the attention blocks, as shown in Fig.7(d).

The FER results of different schemes under a single module are shown in Table III. Scheme 1 obtained the lowest recognition result of 79.35%. We believe that although ROI Align uses bilinear interpolation to pool feature maps to the same size, this process changes the feature value and causes the possibility of important information loss. Scheme 2 obtains a recognition result of 84.65%, which is 0.62% lower than scheme 3. It may be that the unreasonable distribution of facial local regions and the redundant information affect performance. Also, scheme 4 obtains the highest recognition result (86.02%), indicating that pose-based division and location-based padding in the proposed method can eliminate redundant information and retain the location information of subregions effectively while ensuring the integrity of feature transmission.

3) AP module: We also evaluate the performance of the AP module. The performance of the AP module is not outstanding in single module recognition, only reaching 0.3%, 0.5%, and 0.5% higher than ResNet-32 on RAF-DB, AffectNet-7, and FED-RO datasets, as shown in Table I. However, when the AP module is combined with the GP module and the LP modules as auxiliary information, FER performance can be effectively improved by 0.45%, 0.68%, and 1.5% on RAF-DB, AffectNet-7, and FED-RO datasets, respectively, as shown in Table II. Experimental results show that the AP module can be used as supplementary information for the GP module and LP module because it pays more attention to the small facial regions with significant emotion correlation obtained based on prior knowledge, to avoid the lack of important information caused by the inaccurate positioning of facial emotion expression regions by the model and effectively improve the robustness of salient feature extraction.

4) Region size L: The AP module extracts facial feature maps of the eye and mouth regions with $128 \times L \times L$ as auxiliary information. To explore the influence of the region size L, we take the FER performance of L as 6, 8, 10, 12, and 14 under the combination of all modules, as shown in Fig.8(a). Results show that when $L = 10$, AMP-Net has the highest recognition result (88.06%). When $L < 10$, we believe the reduction of the region size will miss some important features; when $L > 10$, although the increase of region size improves the integrity of auxiliary information, more information will lead to an overload of the network and reduce the degradation of identification performance.

5) Weight λ: To explore the influence of the loss function weight λ on AMP-Net, different λ from 0 to 1 are selected. The FER results are shown in Fig.8(b), and indicate that when $\lambda = 0.5$, AMP-Net achieves the highest FER performance, and the global perception features have the same weight as the local and attention perception features, indicating that the complementary features between global, local, and attention facial features can ensure effective information extraction.

D. Comparison with the state-of-the-art methods

In this section, we compare the proposed method's best results to several state-of-the-art methods on RAF-DB, AffectNet, SFEW 2.0, FER-2013, and FED-RO real-world

TABLE IV
COMPARISON TO THE STATE-OF-THE-ART RESULTS ON RAF-DB DATASET * USING RESNET-18 AS BACKBONE

Method	Backbone	Year	Acc.
ACNN [20]	VGG-16	2019	85.07
RAN [39]	ResNet-18	2020	86.9
SCN [55]	ResNet-18	2020	88.14
LBAN-IL [56]	ResNet-18	2021	85.89
MA-Net [18]	ResNet-18	2021	88.40
DACL [13]	ResNet-18	2021	87.78
PASM [57]	VGG-16	2021	87.5
PASM [57]	ResNet-18	2021	87.18
PASM [57]	ResNet-34	2021	88.68
AMP-Net*	ResNet-18	-	88.84
AMP-Net	ResNet-34	-	**89.25**

TABLE V
COMPARISON TO THE STATE-OF-THE-ART RESULTS ON AFFECTNET DATASET * USING RESNET-18 AS BACKBONE

Method	Backbone	Year	Acc.	
			7 classes	8 classes
ACNN [20]	VGG-16	2019	58.78	/
BOVW [58]	VGG-13	2019	63.31	59.58
SCN [55]	ResNet-18	2020	/	60.23
RAN [39]	ResNet-18	2020	/	59.5
OAENet [19]	Manually	2021	58.7	/
MA-Net [18]	ResNet-18	2021	64.53	60.29
Triplet loss [12]	ResNet-18	2021	/	60.12
LAENet-SA [59]	ResNet-18	2021	64.09	61.22
AMP-Net*	ResNet-18	-	64.32	61.39
AMP-Net	ResNet-34	-	**64.54**	**61.74**

TABLE VI
COMPARISON TO THE STATE-OF-THE-ART RESULTS ON SFEW 2.0 DATASET

Method	Pretrained Dataset	Year	Acc.
ACNN [20]	AffectNet	2019	52.59
RAN(ResNet18) [39]	MS-Celeb-1 M	2020	54.19
RAN(VGG18) [39]	MS-Celeb-1 M	2020	56.40
LBAN-IL [56]	ResNet-18	2021	55.28
AHA [60]	ImageNet	2021	58.89
DMUE [61]	MS-Celeb-1 M	2021	58.34
MA-Net [18]	FER-2013	2021	59.4
AMP-Net	RAF-DB	-	**61.17**

TABLE VII
COMPARISON TO THE STATE-OF-THE-ART RESULTS ON FER-2013 DATASET

Method	Backbone	Year	Acc.
PAT-VGG-F [62]	VGG-16	2018	72.16
PAT-ResNet [62]	ResNet-34	2018	72.00
Soft Label [63]	VGG-16	2019	73.73
PASM [57]	VGG-16	2021	72.73
PASM [57]	ResNet-34	2021	73.56
AHA [60]	ResNet-18	2021	73.84
LBAN-IL [56]	Manually	2021	73.11
AMP-Net	ResNet-34	-	**74.48**

datasets as well as Occlusion-AffectNet, Occlusion-RAF-DB, Occlusion-FERPlus, Pose-AffectNet, Pose-RAF-DB, and Pose-FERPlus occlusion and variant pose datasets. All benchmark results are reported in the literature.

1) Real-world datasets

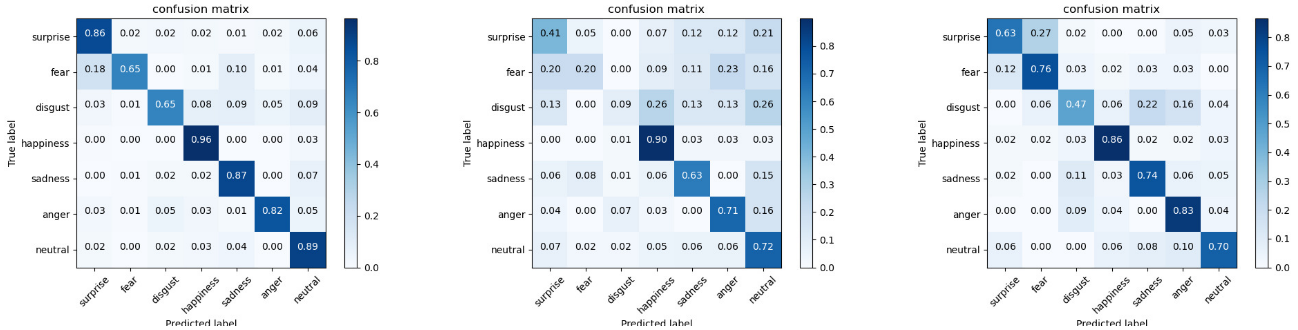
a) Comparison with RAF-DB: Table IV compares the proposed method and the state-of-the-art methods in RAF-DB with seven emotion categories of happiness, neutral, surprise, fear, anger, disgust, and sadness. AMP-Net achieves the highest FER result (89.25%) under the pretraining in the VGGFace2 face dataset on the RAF-DB dataset. In the confusion matrix shown in Fig.9(a), fear and disgust experience high recognition difficulties, and 18% of the fear dataset is incorrectly identified as surprise. The reason for these results may be that both fear and surprise exhibit high confusion in recognition determination and in facial muscle movement [53], [54]. In addition, to fairly compare with previous studies, we also test AMP-Net on ResNet-18 [42] as the backbone network, and the results also outperformed existing methods.

b) Comparison with AffectNet: AffectNet is currently the largest facial expression dataset. Due to the complexity and diversity of facial images on AffectNet, it is difficult to recognise. To fully verify the effectiveness of the proposed method, we select 7 emotion categories and 8 categories with ‘contempt’ added to conduct the experiments. The results under different backbones are shown in Table V. For 7 types of emotion categories, the proposed method performs 0.52% better than the highest known LAENet-SA [59] method (61.22%), although the recognition results of the proposed method are similar to MA-Net [18], and 1.45% higher than MA-Net under 8 types of emotion categories.

c) Comparison with SFEW 2.0: Table VI compares the proposed method and the state-of-the-art methods using the SFEW 2.0 dataset. AMP-Net achieves the highest FER result (61.17%), and MA-Net [18] produces the result with the highest known recognition accuracy (59.40%) by performing multiscale FER through the global region and evenly distributed feature maps of the local region. The proposed method achieves 61.17% in SFEW 2.0, which highlights the FER robustness of AMP-Net. Fig.9(b). shows the recognition confusion matrix. Fear and disgust achieve low recognition results primarily due to the scarcity of emotion images in SFEW 2.0 and the ease of confusing fear and disgust.

d) Comparison with FER-2013: The FER-2013 dataset is a set of greyscale images of average quality that were collected via network search, and problems such as blurry images and missing labels make recognition difficult. In the experiment, we do not apply the data enhancement method of random changes in brightness, contrast, and saturation, and results are shown in Table VII. The proposed method achieves the highest FER result (74.48%), demonstrating that AMP-Net can capture effective robust facial features in different situations.

e) Comparison with FED-RO: The FED-RO dataset contains specially collected face images with occlusion. In the experiment, using the same settings as RAN [39], training on the training set of the AffectNet-7 and RAF-DB datasets, and testing on the FER-RO dataset, the results are shown in Table VIII. The proposed method achieves the highest recognition result (71.75%) among known FER methods. Experimental results show that AMP-Net can adapt to the problem of facial occlusion more effectively and has a higher generalisability. In addition, Fig.9(c) shows the confusion matrix of the proposed method for FED-RO. AMP-Net achieves better FER performance for most emotion categories under occlusion, and the incorrect recognition of FED-RO is primarily due to the high



(a) RAF-DB (b) SFEW 2.0 (c) FED-RO

Fig. 9. Confusion matrix of the RAF-DB, SFEW 2.0, and FED-RO dataset. (a) Confusion matrix of the RAF-DB, (b) Confusion matrix of the SFEW 2.0, (c) Confusion matrix of the FED-RO.

TABLE VIII
COMPARISON TO THE STATE-OF-THE-ART RESULTS ON FED-RO DATASET

Method	Year	Acc. (%)
ACNN [20]	2019	66.5
SPWFA-SE [21]	2020	67.25
LAENet-SA [59]	2021	68.25
IDFL [38]	2021	67.25
MA-Net [18]	2021	70.00
AMP-Net	-	**71.75**

confusion of fear, surprise, and disgust.

2) Occlusion and variant-pose datasets

To investigate the robustness of AMP-Net under occlusion and variant poses, we compare its best results with the occlusion and variant-pose datasets, including Occlusion-RAF-DB, Occlusion-AffectNet, Occlusion-FERPlus, Pose-AffectNet, Pose-RAF-DB, Pose-AffectNet, and Pose-FERPlus.

Table IX, X, and XI compares the proposed method and state-of-the-art methods in the RAF-DB, AffectNet and FERPlus datasets for facial occlusion and variant poses. AMP-Net achieves superior performance compared to the benchmark method RAN [39] and achieves the highest FER performance. For facial occlusion, the proposed method achieves outstanding recognition performance (85.28%, 64.27%, and 85.44%) in the RAF-DB, AffectNet and FERPlus datasets. In particular, AMP-Net surpasses the ASF [64] method with the highest known accuracy by 1.33%, 1.29% and 0.65%, which further demonstrates the high robustness of AMP-Net for extracting effective emotion features under facial occlusion. For variant poses, the proposed method also achieves the highest recognition performance in the RAF-DB, AffectNet and FERPlus datasets under Pose≥ 30° and pose≥ 45°. Table IX, X, and XI show that the FER results of Pose≥ 45° are all lower than the FER results of Pose≥ 30°, indicating that the larger the pose variation is, the more difficult it is to extract effective information. The LP module in AMP-Net uses an adaptive approach for different facial poses to acquire finer subregion facial features. Results demonstrate the high robustness of AMP-Net for variant poses, particularly for the FERPlus dataset. The proposed method performs 5.29% and 7.17% better than RAN under Pose≥ 30° and Pose≥ 45°, respectively.

TABLE IX
COMPARISON TO THE STATE-OF-THE-ART RESULTS ON OCCLUSION-RAF-DB, POSE-RAF-DB DATASET

Method	Occlusion	Pose(≥30°)	Pose(≥45°)
RAN [39]	82.72	86.74	85.2
MA-Net [18]	83.65	89.66	87.99
ASF [64]	83.95	87.89	88.35
AMP-Net	**85.28**	**89.75**	**89.25**

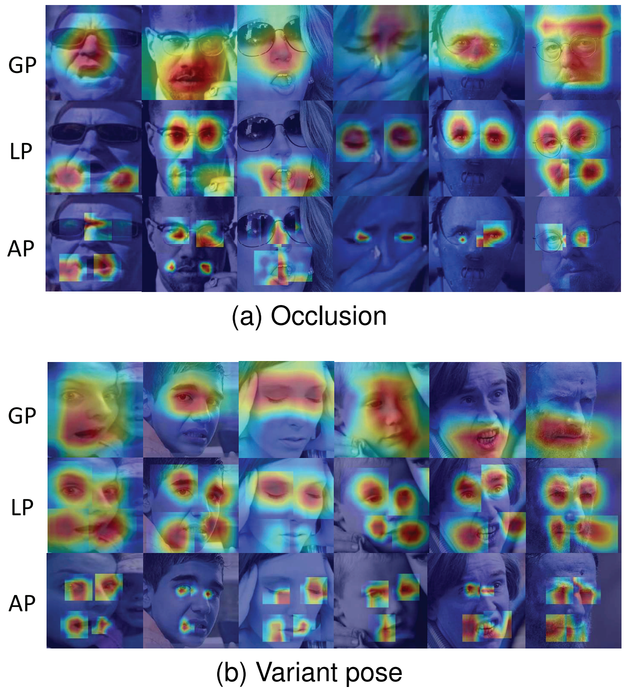
TABLE X
COMPARISON TO THE STATE-OF-THE-ART RESULTS ON OCCLUSION-AFFECTNET, POSE- AFFECTNET DATASET

Method	Occlusion	Pose(≥30°)	Pose(≥45°)
RAN [39]	58.5	53.9	53.19
MA-Net [18]	59.59	57.51	57.78
ASF [64]	62.98	60.61	61
AMP-Net	**64.27**	**61.37**	**61.16**

TABLE XI
COMPARISON TO THE STATE-OF-THE-ART RESULTS ON OCCLUSION-FERPLUS, POSE- FERPLUS DATASET

Method	Occlusion	Pose(≥30°)	Pose(≥45°)
RAN [39]	83.63	82.23	80.4
ASF [64]	84.79	88.29	87.2
AMP-Net	**85.44**	**88.52**	**87.57**

In addition, to investigate the performance of AMP-Net in more detail, we use gradient weighted class activation mapping (Grad-CAM) [54] to visualise attention maps of AMP-Net under occlusion and variant poses, As shown in Fig.10, the attention of the GP module, LP module and AP module to different facial regions is shown, and dark red indicates areas of high concern. Face occlusion and pose variation markedly change facial visual appearances, as shown in Fig.10. For face occlusion, the GP module can focus on unoccluded facial areas from a global perspective, which indicates high robustness to occlusion (see Fig.10(a)). The LP module adaptively divides the face into four finer parts based on the head pose, as shown in Fig.10(a). This module eliminates ineffective occluded subregions, such as hands, masks, and sunglasses, and pays more attention to unoccluded eyes and mouth areas. Due to the small occlusion area of glasses, the LP module can adaptively focus on the unobstructed eye region to obtain



(a) Occlusion

(b) Variant pose

Fig. 10. Attention maps of several occlusion and pose variation facial images on FED-RO, AffectNet, RAF-DB datasets. GP denotes GP module, LP denotes LP module, AP denotes AP module. (a) Attention maps of occlusion facial images, (b) Attention maps of variant pose facial images.

robust emotional features when people wear glasses. As a supplementary module of AMP-Net, AMP-Net pays more attention to small salient areas of the eyes and mouth with high emotional correlation. Fig.10(a) shows the adaptability of the AP module under facial occlusion. The LP module can allocate facial organs such as the eyes and mouth in variant poses to more refined subregions, and focus on the eyes and mouth regions that are similar to the human visual attention mechanism. Fig.10(b) also shows the efficient adaptability of the proposed method under variant poses. Therefore, these results demonstrate the high robustness of AMP-Net for facial occlusion and variant poses.

E. Error analysis

Fig.11 shows example images with occlusion and variant poses in the FED-RO and AffectNet datasets where AMP-Net failed to predict the correct expression categories. Although AMP-Net is robust to different facial conditions, incomplete facial information is caused by occlusion and variant poses. The low emotional expression intensity and the similar facial action units (AUs) of the unoccluded regions are highly likely to cause recognition errors. As shown in Fig. 11, during mouth occlusion and pose$\geq$ 30°, *surprise* is wrongly identified as *fear*, *anger* is wrongly identified as *disgust*, etc., because *surprise* and *fear* have similar inner brow raisers and upper lid raisers (AU1, 5), and *anger*, *disgust* and *sadness* have similar brow lowers and Lip corner depressors (AU4, 15). The solution to these problems is to supplement multimodal explicit behaviour information, such as body or language information, to enhance emotional differences.

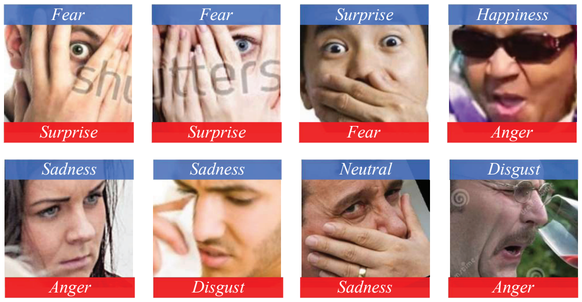


Fig. 11. Some example images with occlusion and variant pose on FED-RO and AffectNet datasets that AMP-Net failed to predict the correct expression categories with. Note that 'blue' represents true labels and 'red' represents prediction labels.

V. Conclusion

In this paper, we propose an adaptive multilayer perceptual attention network (AMP-NET) that is inspired by the facial attributes and human visual perception mechanism to acquire multilevel facial emotional features from coarse to fine to improve robustness under occlusion and variant poses. We design three modules to obtain facial information from different perception domains to ensure that the model pays more attention to facial regions with substantial emotional correlation and robustness to real-world facial emotion data. The final ablation experiment and comparison with the existing methods show that the proposed method can effectively eliminate invalid information under occlusion and exhibits high robustness to facial occlusion and variant poses. In future work, we plan to investigate the construction of a multimodal emotion recognition model based on federated learning to improve the model generalisability and recognition accuracy in real-world scenarios to ensure user privacy.

References

[1] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 3030–3043, 2017.

[2] F. Noroozi, C. A. Corneanu, D. Kamińska, T. Sapiński, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE transactions on affective computing*, vol. 12, no. 2, pp. 505–523, 2018.

[3] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, 2020.

[4] A. Samara, L. Galway, R. Bond, and H. Wang, "Affective state detection via facial expression analysis within a human–computer interaction context," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 6, pp. 2175–2184, 2019.

[5] L. He, C. Guo, P. Tiwari, H. M. Pandey, and W. Dang, "Intelligent system for depression scale estimation with facial expressions and case study in industrial intelligence," *International Journal of Intelligent Systems*, 2021.

[6] 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 *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE, 2010, pp. 94–101.

[7] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proceedings Third IEEE international conference on automatic face and gesture recognition*. IEEE, 1998, pp. 200–205.

[8] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE international conference on multimedia and Expo*. IEEE, 2005, pp. 5–pp.

[9] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2018.

[10] A. Mollahosseini, B. Hasani, and 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, 2017.

[11] Y. Li, Y. Gao, B. Chen, Z. Zhang, G. Lu, and D. Zhang, "Self-supervised exclusive-inclusive interactive learning for multi-label facial expression recognition in the wild," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[12] W. Xie, H. Wu, Y. Tian, M. Bai, and L. Shen, "Triplet loss with multistage outlier suppression and class-pair margins for facial expression recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[13] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2402–2411.

[14] X. Zhang, F. Zhanga, and C. Xu, "Joint expression synthesis and representation learning for facial expression recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[15] L. Yang, Y. Tian, Y. Song, N. Yang, K. Ma, and L. Xie, "A novel feature separation model exchange-gan for facial expression recognition," *Knowledge-Based Systems*, vol. 204, p. 106217, 2020.

[16] C. Chen and R. E. Jack, "Discovering cultural differences (and similarities) in facial expressions of emotion," *Current opinion in psychology*, vol. 17, pp. 61–66, 2017.

[17] Y. Xia, W. Zheng, Y. Wang, H. Yu, J. Dong, and F.-Y. Wang, "Local and global perception generative adversarial network for facial expression synthesis," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[18] Z. Zhao, Q. Liu, and S. Wang, "Learning deep global multi-scale and local attention features for facial expression recognition in the wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 6544–6556, 2021.

[19] Z. Wang, F. Zeng, S. Liu, and B. Zeng, "Oaenet: Oriented attention ensemble for accurate facial expression recognition," *Pattern Recognition*, vol. 112, p. 107694, 2021.

[20] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2018.

[21] Y. Li, G. Lu, J. Li, Z. Zhang, and D. Zhang, "Facial expression recognition in the wild using multi-level features and attention mechanisms," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2020.

[22] D. Roberson, M. Kikutani, P. Döge, L. Whitaker, and A. Majid, "Shades of emotion: What the addition of sunglasses or masks to faces reveals about the development of facial expression processing," *Cognition*, vol. 125, no. 2, pp. 195–206, 2012.

[23] J. Duncan, F. Gosselin, C. Cobarro, G. Dugas, C. Blais, and D. Fiset, "Orientations for the successful categorization of facial expressions and their link with facial features," *Journal of vision*, vol. 17, no. 14, pp. 7–7, 2017.

[24] S. C. Widen, A. M. Christy, K. Hewett, and J. A. Russell, "Do proposed facial expressions of contempt, shame, embarrassment, and compassion communicate the predicted emotion?" *Cognition & Emotion*, vol. 25, no. 5, pp. 898–906, 2011.

[25] Y. Wang, Z. Zhu, B. Chen, and F. Fang, "Perceptual learning and recognition confusion reveal the underlying relationships among the six basic emotions," *Cognition and Emotion*, vol. 33, no. 4, pp. 754–767, 2019.

[26] L. A. Halliday, "Emotion detection: can perceivers identify an emotion from limited information?" 2008.

[27] X. Yan, T. J. Andrews, and A. W. Young, "Cultural similarities and differences in perceiving and recognizing facial expressions of basic emotions." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 42, no. 3, p. 423, 2016.

[28] Y. Busin, K. Lukasova, M. K. Asthana, and E. C. Macedo, "Hemiface differences in visual exploration patterns when judging the authenticity of facial expressions," *Frontiers in psychology*, vol. 8, p. 2332, 2018.

[29] M. E. Nicholls, B. E. Ellis, J. G. Clement, and M. Yoshino, "Detecting hemifacial asymmetries in emotional expression with three–dimensional computerized image analysis," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 271, no. 1540, pp. 663–668, 2004.

[30] U. Dimberg and M. Petterson, "Facial reactions to happy and angry facial expressions: Evidence for right hemisphere dominance," *Psychophysiology*, vol. 37, no. 5, pp. 693–696, 2000.

[31] R. E. Jack, C. Blais, C. Scheepers, P. G. Schyns, and R. Caldara, "Cultural confusions show that facial expressions are not universal," *Current biology*, vol. 19, no. 18, pp. 1543–1548, 2009.

[32] C. A. Hutcherson and J. J. Gross, "The moral emotions: A social–functionalist account of anger, disgust, and contempt." *Journal of personality and social psychology*, vol. 100, no. 4, p. 719, 2011.

[33] A. Brooke and N. Harrison, "Neuroimaging and emotion," in *Stress: Concepts, cognition, emotion, and behavior*. Elsevier, 2016, pp. 251–259.

[34] H. Eisenbarth and G. W. Alpers, "Happy mouth and sad eyes: scanning emotional facial expressions." *Emotion*, vol. 11, no. 4, p. 860, 2011.

[35] M. Schurgin, J. Nelson, S. Iida, H. Ohira, J. Chiao, and S. Franconeri, "Eye movements during emotion recognition in faces," *Journal of vision*, vol. 14, no. 13, pp. 14–14, 2014.

[36] L. Zhang, K. Mistry, M. Jiang, S. C. Neoh, and M. A. Hossain, "Adaptive facial point detection and emotion recognition for a humanoid robot," *Computer Vision and Image Understanding*, vol. 140, pp. 93–114, 2015.

[37] J. Y. R. Cornejo, H. Pedrini, and F. Flórez-Revuelta, "Facial expression recognition with occlusions based on geometric representation," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2015, pp. 263–270.

[38] Y. Li, Y. Lu, B. Chen, Z. Zhang, J. Li, G. Lu, and D. Zhang, "Learning informative and discriminative features for facial expression recognition in the wild," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[39] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.

[40] L. Zhang, D. Tjondronegoro, and V. Chandran, "Random gabor based templates for facial expression recognition in images with facial occlusion," *Neurocomputing*, vol. 145, pp. 451–464, 2014.

[41] H. Ding, P. Zhou, and R. Chellappa, "Occlusion-adaptive deep network for robust facial expression recognition," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2020, pp. 1–9.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[43] Y. Lee, J.-w. Hwang, S. Lee, Y. Bae, and J. Park, "An energy and gpu-computation efficient backbone network for real-time object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[44] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "Densenet: Implementing efficient convnet descriptor pyramids," *arXiv preprint arXiv:1404.1869*, 2014.

[45] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[46] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5203–5212.

[47] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[48] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE multimedia*, vol. 19, no. 03, pp. 34–41, 2012.

[49] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International conference on neural information processing*. Springer, 2013, pp. 117–124.

[50] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[51] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[52] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[53] S. Namba, R. S. Kabir, M. Miyatani, and T. Nakao, "Spontaneous facial actions map onto emotional experiences in a non-social context: toward a component-based approach," *Frontiers in Psychology*, vol. 8, p. 633, 2017.

[54] S. Du and A. M. Martinez, "The resolution of facial expressions of emotion," *Journal of Vision*, vol. 11, no. 13, pp. 24–24, 2011.

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

[56] H. Li, N. Wang, Y. Yu, X. Yang, and X. Gao, "Lban-il: A novel method of high discriminative representation for facial expression recognition," *Neurocomputing*, vol. 432, pp. 159–169, 2021.

[57] P. Liu, Y. Lin, Z. Meng, L. Lu, W. Deng, J. T. Zhou, and Y. Yang, "Point adversarial self-mining: A simple method for facial expression recognition," *IEEE Transactions on Cybernetics*, 2021.

[58] M.-I. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64 827–64 836, 2019.

[59] C. Wang, J. Xue, K. Lu, and Y. Yan, "Light attention embedding for facial expression recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[60] J. Weng, Y. Yang, Z. Tan, and Z. Lei, "Attentive hybrid feature with two-step fusion for facial expression recognition," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 6410–6416.

[61] J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, and T. Mei, "Dive into ambiguity: latent distribution mining and pairwise uncertainty estimation for facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6248–6257.

[62] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Probabilistic attribute tree in convolutional neural networks for facial expression recognition," *arXiv preprint arXiv:1812.07067*, 2018.

[63] Y. Gan, J. Chen, and L. Xu, "Facial expression recognition boosted by soft label with a diverse ensemble," *Pattern Recognition Letters*, vol. 125, pp. 105–112, 2019.

[64] F. Ma, B. Sun, and S. Li, "Robust facial expression recognition with convolutional visual transformers," *arXiv preprint arXiv:2103.16854*, 2021.

Hanwei Liu received the B.S. degree from Yantai University, Yantai, China, in 2017, the M.S. Degree from Nangjing University of Science and Technology, Nanjing, China, in 2020. He is currently pursuing the Ph.D. degree in control theory and control engineering with the Department of Control Science and Engineering, Tongji University, Shanghai, China. His research interests include image processing and affective computing.

Huiling Cai received the B.S. degree from Jiangnan University, Wuxi, China, in 2017. She is currently pursuing the Ph.D. degree in control theory and control engineering with the Department of Control Science and Engineering, Tongji University, Shanghai, China. Her research interests include physiological information processing and affective computing.

Qingcheng Lin received the B.S. degree from Northeastern University, Shenyang, China, in 2018. She is currently pursuing the Ph.D. degree in control theory and control engineering with the Department of Control Science and Engineering, Tongji University, Shanghai, China. Her research interests include image processing and affective computing.

Xuefeng Li (M'17) graduated from Shenyang Institute of Engineering, China, in 1999. He received his M.E. and D.E. degrees from Fukuoka Institute of Technology, Fukuoka, Japan, in 2004 and 2007, respectively. From 2007 to 2013, he was a postdoctoral and assistant researcher at Waseda University, Kitakyushu, Japan. Since 2013, he has been an associate professor at Tongji University. His research interests include intellisense and information processing.

Hui Xiao received the B.S., M.S. and Ph.D. degrees from Tongji University, Shanghai, China, in 1992, 1998 and 2007, respectively. From 1992 to 1997, she was a lecturer at Tongji University. From 1997 to 2011, she was an associate professor at Tongji University. Since 2011, she has been a professor at Tongji University. Her research interests include intelligent control, image processing, affective analysis, and application novel techniques.