FERMixNet: An Occlusion Robust Facial Expression Recognition Model with Facial Mixing Augmentation and Mid-Level Representation Learning

Yansong Huang, Junjie Peng, *Member, IEEE,* Wenqiang Zhang, Tong Zhao, Gan Chen, Shuhua Tan, Fen Yi and Lu Wang

Abstract—Facial expressions can provide a better understanding of people's mental status and attitudes towards specific things. However, facial occlusion in real world is an unfavorable phenomenon that greatly affects the performance of facial expression recognition models. Recent works addressing the occlusion problem have primarily relied on attention mechanisms or occlusion discarding methods that focus on non-occluded regions of the face. However, these methods have not achieved a good balance between occlusion robustness and model efficiency. In this paper, we propose a simple and efficient model, called FERMixNet, for occluded facial expression recognition. The model incorporates a novel facial mixing augmentation strategy (FERMix) that generates new training samples by simulating real-world facial occlusion and preserving high expression-related semantic information. By co-training the original and newly generated samples, the model's occlusion robustness is improved without increasing its complexity during inference. Additionally, to further enhance the model's occlusion robustness, we include mid-level representation learning in the network to learn the discriminative non-occluded local features of the samples with low computational cost. Extensive experiments on four public facial occlusion datasets: Occlusion-RAF-DB, Occlusion-FERPlus and FED-RO show that the proposed model achieves state-of-the-art results which demonstrates the good robustness of our method for occluded facial expression recognition. Meanwhile, the proposed model also achieves state-of-the-art results on the in-the-wild facial expression datasets RAF-DB, AffectNet-8, and AffectNet-7. It proves that the proposed model has good application prospects in real world.

Index Terms—Facial Expression Recognition, Facial Occlusion, FERMixNet, Facial Mixing Augmentation, Mid-level Representation Learning.

1 Introduction

FACIAL expressions are one of the most prevalent and important signals for conveying human emotions and intentions [1], [2]. It is of great importance to correctly recognize facial expressions. Because of this, facial expression recognition (FER) has become a fundamental task in the field of computer vision, and has been widely used in applications such as online education, healthcare, home escort, product recommendation, video recommendation and online monitoring, to obtain user preferences and psychological states to achieve humanized human-computer interaction.

With the development of deep learning, a large number of deep learning-based facial expression recognition methods has emerged. Among these, the topic of occluded facial expression recognition is more challenging and it has received increasing attention in recent years. This is attributed to the fact that the face images captured by machines in the real world are not always complete, and occlusions are a common phenomenon. Large in-the-wild facial expression datasets such as RAF-DB [3], FERPlus [4], and AffectNet [5] contain a significant amount of occluded data. By observing these data, we categorize occlusion types into three classes based on the location of occlusions, as shown in Fig. 1. Facial occlusions may occur due to various factors such as mobile phones, glasses, masks, hands, hair, etc., leading to the loss of crucial facial information and a significant reduction in

(a) Upper and lower (b) Left and right (c) Corner

Fig. 1: Occlusion types: (a) upper and lower occlusions; (b) left and right occlusions; (c) corner occlusions.

- *Y. Huang, T. Zhao, G. Chen and L. Wang are with School of Computer Engineering and Science, Shanghai University, Shanghai, China. E-mail: huangyansong@shu.edu.cn.*
- *J. Peng is with the School of Computer Engineering and Science & the Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai, China. E-mail: jjie.peng@shu.edu.cn.*
- *W. Zhang is with the Academy for Engineering and Technology, and School of Computer Science and Technology, Fudan University, Shanghai, China. E-mail: wqzhang@fudan.edu.cn.*
- *S. Tan and F. Yi are with YTO Express Co., Ltd. and the national logistics engineering laboratory, Shanghai, China.*

Corresponding author: Junjie Peng(email:jjie.peng@shu.edu.cn).

the performance of facial expression recognition systems. Recently some studies [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]are devoted to improving the occlusion robustness of facial expression recognition systems. Generally, the solution ideas of these studies for this problem are as follows.

Idea 1. Face complementation [6], [7], [8]. This idea is to use a generative approach to complement the face of the occluded part and then feed the generative complemented face into the facial expression recognition network for classification. However, such methods rely on the generative model to learn good expression features, which is challenge to train. Furthermore, using human-simulated occlusions instead of real-world occlusion scenarios makes existing methods difficult to generalize to realistic scenarios.

Idea 2. Occlusion discarding [9], [10], [11]. The idea is to address occlusion by masking the occluded region, thereby minimizing its impact on model performance. However, implementing this idea typically necessitates additional prior information and a deep learning model to identify the occlusion's location, which significantly increases the model's complexity and computational overhead, and is not practical in real-world applications.

Idea 3. Enhancing non-occluded local feature representation [12], [13], [14], [15], [16]. Psychological studies indicate that human can effectively exploit both local regions and holistic faces to perceive the semantics delivered through incomplete faces [17]. When some parts of the face are occluded, humans can judge expressions based on other local non-occluded regions. Based on these studies, many approaches attempt to enhance the ability of face local regions to represent expression features. The current prevailing approaches of this idea are using attention mechanism to make the model focus on non-occluded local regions.

Idea 4. Expand the dataset and data augmentation. Adding more occluded facial expression images in the dataset is one of the most direct and effective ways, but the difficulty lies in the fact that dataset collection is time-consuming and labor-intensive. Meanwhile there is no large occluded facial expression dataset directly available for training so far. Thus, the most effective approach to implementing this idea is to use data augmentation strategies, such as Cutout [18], which involves adding simulated occlusion samples to the training data, thereby enhancing the occlusion robustness of the model. The key benefit of this approach is that it leads to improved occlusion robustness without incurring additional computational cost or model complexity.

The current mainstream research on occluded facial expression recognition focuses on enhancing non-occluded local region representation using attention mechanisms. ACNN [12] and RAN [13] improve model robustness by manually dividing face local regions and enhancing the feature representation of non-occluded local regions with attention mechanisms. However, the above methods require manual division of local regions in the inference phase, which greatly increases the computational complexity. VTFF [14] uses global and local attention mechanisms to enhance feature representation, but requires the extraction of face LBP features [19] and introduces the Transformer module [20], significantly increasing model complexity and training difficulty. In contrast, EfficientFace [15] uses a spatial-channel attention mechanism to focus on non-occluded local regions and employs a lighter backbone to reduce model complexity. Additionally, label distribution is used to improve the model's generalization ability. EfficientFace is a highly efficient network, but the use of a lightweight model may limit its performance to some extent.

Several current methods based on occlusion discarding have achieved good results. MAPNet [10] removes background and occlusion information using facial landmarks. However, facial landmark detection techniques are also limited by occlusion conditions, making it difficult to ensure their effectiveness under such conditions. MViT [9] uses a transformer-based mask generation network to generate occlusion masks, while Co-completion [11] uses a segmentation network and an additional guidance network to generate occlusion masks, reducing the impact of occlusion. However, these methods require additional mask generation networks, which greatly increases the model complexity, and it is difficult to ensure the quality of mask generation without good data support.

In order to keep high efficiency while enhancing the occlusion robustness of the model, we design a simple and efficient occlusion robustness facial expression recognition model called FERMixNet. The proposed model leverages two previously mentioned ideas: data augmentation and enhancing non-occluded local feature representation. Specifically, we design a novel FERMix data augmentation strategy, which generates mixed images that simulate three common occlusion cases in real-world scenarios while preserving the high expression-related semantic information. The mixed images are used as additional training samples and co-trained with the corresponding original images to enhance the model's occlusion robustness. During inference, the mixed images are not utilized, hence avoiding additional computational complexity. Additionally, to enhance non-occluded local feature representation, we incorporate a mid-level representation learning branch into the traditional deep convolutional network, which filters out useless occlusion information and highlights the most discriminative non-occluded local features at low computational cost. Our experiments demonstrate that FERMixNet significantly improves facial expression recognition performance in occlusion conditions and performs consistently well in realistic scenarios.

The contributions of our work are summarized as follows.

(1) We propose FERMixNet for occluded facial expression recognition. By using the facial mixing augmentation and mid-level representations, the model improves the occlusion robustness of facial expression recognition.

(2) To enhance the ability of face regions to represent expression features, we propose a facial mixing augmentation strategy FERMix for occluded facial expressions recognition. The method generates new training samples for training by simulating real-world occlusion situations while mixing the high expression-related semantically informative local regions of two face images.

(3) To learn the most discriminative non-occluded local features and further improve the occlusion robustness of the model, we introduce a mid-level representation branch

in the backbone model, while it just needs few costs.

(4) Extensive experiments are performed to evaluate the validity of the proposed model in occlusion facial expression recognition. The proposed model achieves the state-of-the-art results on three facial occlusion datasets: Occlusion-RAF-DB, Occlusion-FERPlus, and FED-RO, with accuracies of 86.67%, 85.95%, and 72.97%, respectively. And it also achieves state-of-the-art results on three in-the-wild facial expression datasets: RAF-DB, AffectNet-8, and AffectNet-7, with accuracies of 91.62%, 63.22%, and 66.40%, respectively.

The rest of the paper is organized as follows. Section 2 describes the related work. Section 3 details the proposed model. Section 4 shows the experimental results and analyzes them. Section 5 concludes the whole paper.

2 RELATED WORK

In this section we mainly present some recent studies about occluded facial expression recognition, cutting-based and mixing-based data augmentation strategy, and mid-level representation learning.

2.1 Occluded Facial Expression Recognition

Occluded facial expression recognition is an important branch of facial expression recognition tasks, and has received increasing attention by many researchers in recent years. Current state-of-the-art methods in occluded facial expression recognition can be categorized into two main categories: methods based on enhancing non-occluded local feature and methods based on occlusion discarding.

Methods based on enhancing non-occluded local feature are mainstream. These methods enhance the representation of non-occluded local features through attention mechanisms. For instance, ACNN [12] manually selects 24 facial landmarks and input the corresponding local region features to an attention network, which automatically learns the weights of the local regions based on the level of occlusion. Similarly, RAN [13] manually crops fixed regions of the face and uses a self-attention model to learn the weight of each region. However, these methods require manual partitioning of local regions during the inference stage, significantly increasing inference complexity. VTFF [14] employs global and local attention mechanisms to enhance feature representation but requires the extraction of facial LBP features [19] as input beforehand. Additionally, it introduces the Transformer module [20], which significantly increases model complexity and training difficulty. Wang et al. [16] proposed a light attention embedding network that leverages spatial attention mechanisms to improve occlusion robustness. In contrast, FERMixNet requires no additional manual operations, being an end-to-end network. The FERMix strategy is only used during the training phase, thus not affecting inference efficiency. Compared to methods using attention mechanisms, our introduced mid-level representation features incur lower costs.

Compared to methods that enhance non-occluded local features, occlusion discarding methods reduce the impact of occluded regions through masks. Some recent studies based on occlusion discarding have achieved promising results. Ju et al. [10] introduced a mask-based attention parallel network (MAPNet), which uses facial landmark detection to obtain expression-related region masks. However, facial landmark detection is also limited under occlusion conditions. Li et al. [9] used a transformer-based mask generation network to produce occlusion masks, while Zhen et al. [11] utilized a U-Net segmentation network and an additional guidance network to generate occlusion masks, reducing the impact of occlusion. However, these methods require additional mask generation networks, greatly increasing model complexity. And it is challenging to ensure the quality of mask generation without good data support. However, FERMixNet is a single-stage network that does not require the use of additional facial landmark detection or mask generation networks, thereby having better stability and faster inference speed.

2.2 Cutting-based and Mixing-based Data Augmentation

Currently, cutting-based and mixing-based data augmentation strategies are widely used in image classification tasks to improve the model's generalization ability. Cutout [18], a representative of the Cutting-based strategy, replaces random local image regions with zeros to simulate occlusion situations. It forces the model to learn to recognize objects even when they are partially occluded, thus improving the occlusion robustness of the model. Mixup [21] is a representative of the Mixing-based strategy, which generates new training samples by linearly combining pairs of images and their corresponding labels. It is effective for dealing with noisy data because it encourages the model to learn a more generalized decision boundary and reduces the impact of noisy examples. As occluded samples can be considered as a type of noisy sample, Mixup is also effective for occluded samples. CutMix [22] can be described as a combination of Cutout and Mixup. Like Cutout, it involves cutting out a rectangular patch from an image during training to simulate occlusions. However, instead of removing the patch completely, CutMix pastes it onto the corresponding location of another image, creating a new training example that contains information from both images. This process is similar to that of Mixup, where pairs of training examples are blended together to create new synthetic examples. In this way, CutMix combines the occlusion modeling capability of Cutout with the data mixing capability of Mixup, providing a more powerful data augmentation strategy for improving the performance of deep learning models. However, applying CutMix to facial expression recognition tasks has limitations. Studies have shown that the eyes and mouth area contain the highest expression-related semantic information, while other regions have relatively limited semantic information. As the selected regions by CutMix are random, it is likely that the mixed regions contain few semantic information, which may result in noise that affects model training. In contrast, the proposed FERMix strategy takes advantage of this prior information to develop a corresponding scheme that mixes regions with the highest semantic information, avoiding the appearance of noisy regions while retaining the advantages of CutMix.

2.3 Mid-level Representation Learning

Many fine-grained recognition studies introduce a mid-level representation learning branch in the model to obtain discriminative local features in samples. Combining the mid-level representation learning branch with the high-level representation learning branch that learns global information can capture subtle differences in samples. To enhance the mid-level representation learning, Wang et al. [23] introduced a cross-channel pooling layer. Zhang et al. [24] directly combined mid-level features and high-level features to build a powerful expert network of fine-grained recognition. Huang et al. [25] performed element-wise swapping for partial features while learning mid-level features, in order to improve the diversity of the features. These mid-level models exhibit the ability to enhance local feature representations and complement high-level feature representations. Moreover, mid-level representations are easy to obtain and the model design is simple and low-cost. Hence, we introduce mid-level representation learning branch in our model to learn discriminative local features of the face images, thereby enhancing the representation of non-occluded local regions of the face.

3 METHODOLOGY

3.1 Framework Overview

To address the occlusion problem in facial expression recognition, we design a simple and efficient model FERMixNet for occluded facial expression recognition. The framework of the model is shown in Fig. 2, which consists of three modules: (1) feature enhancement module; (2) feature representation module; (3) model decision module.

The feature enhancement module is reflected in the input of the model which contains the main image and mixed image. The main image is the original training sample, and the mixed image is the new training sample generated by FERMix strategy, a novel facial mixing augmentation strategy. The mixed image simulates the face occlusion in real scenes and contains the local information of the main image. By co-training the main image and the mixed image, the feature representation of the local face information is enhanced and the occlusion robustness of the model is improved. The designed FERMix is efficient, as it only operates in the training phase and does not increase the inference time of the model.

The feature representation module is used to extract high-level and mid-level feature representations of the input samples. To enhance the occlusion robustness of the model, we incorporate a mid-level representation learning branch with the backbone model. The mid-level representation captures discriminative local features of the input samples, allowing for the filtering of useless occlusion information and reduction of the negative influence of occlusion regions. Combining the mid-level representation with the high-level representation learned by the backbone model significantly improves the occlusion robustness performance.

The model decision module contains the calculation of the loss function and the representation of the final decision.

The backbone model we have chosen is ResNet [26], which contains the $conv1$, $conv2\_x$, $conv3\_x$, $conv4\_x$, $conv5\_x$, and fully connected (FC) layers.

3.2 Feature Enhancement

3.2.1 FERMix

The design principle of FERMix is based on two important conditions. Firstly, the generated samples should simulate realistic occlusion situations. Secondly, the face region for mixing needs to contain high expression-related semantic information. To satisfy these conditions, FERMix is designed in three different patterns as illustrated in Fig. 3, including horizontal equal mixing, vertical equal mixing, and corner mixing. Each of these patterns corresponds to a realistic scenario of facial occlusion, such as glasses or mask for upper and lower facial occlusion, answering a phone call for left and right facial occlusion, and single eye patch for corner occlusion. In order to retain high expression-related semantic information, the three patterns preserve the eyes and mouth regions as much as possible, as previous studies [27], [28] have shown that these are the most important parts for facial expression recognition.

The facial expression samples after face alignment have some unique characteristics. For instance, the human eyes are typically positioned in the upper part of the image, while the mouth is located in the lower part. Additionally, most face images demonstrate left-right symmetry due to the symmetrical nature of the human face. Based on these characteristics, we implement FERMix, as depicted in Fig. 3. The main images are the samples in a normal training batch, while the paired images are acquired by random shuffling, as shown in Fig. 4.

Let W and H denote the width and the height of input images. The selected mixed region $\boldsymbol{B} = (r_x, r_y, r_w, r_h)$, where (r_x, r_y) denotes the coordinates of the upper left corner of the mixed region, and (r_w, r_h) denotes the width and height of the mixed region.

(a) **Horizontal Equal Mixing**. The eyes in a face image are generally located in the upper part and the mouth is generally located in the lower part, so it is straightforward to divide the eyes region and the mouth region by slicing the image in half horizontally. The mixed regions are defined as Eq. (1) shows.

$$\boldsymbol{B}_{a_{up}} = \left(0, 0, W, \frac{H}{2}\right), \quad \boldsymbol{B}_{a_{low}} = \left(0, \frac{H}{2}, W, \frac{H}{2}\right) \quad (1)$$

$\boldsymbol{B}_{a_{up}}$ represents that the mixed region is the upper half of the image, i.e., the area of both eyes. $\boldsymbol{B}_{a_{low}}$ indicates that the mixed region is the lower half of the image, i.e., the mouth area. This mixing pattern is designed to simulate scenarios where the upper or lower part of the face is occluded, such as when wearing a mask that covers the lower face or sunglasses that obscure the upper part. The generated samples by horizontal equal mixing allows the model to focus more on the upper or lower region of the face, increasing the robustness of the model to upper and lower occlusions.

(b) **Vertical Equal Mixing**. As the human face is generally symmetrical, the left and right facial regions can be obtained by dividing the facial image vertically into two halves. The mixed regions are defined as Eq. (2) shows.

$$\boldsymbol{B}_{b_{left}} = \left(0, 0, \frac{W}{2}, H\right), \quad \boldsymbol{B}_{b_{right}} = \left(\frac{W}{2}, 0, \frac{W}{2}, H\right) \quad (2)$$

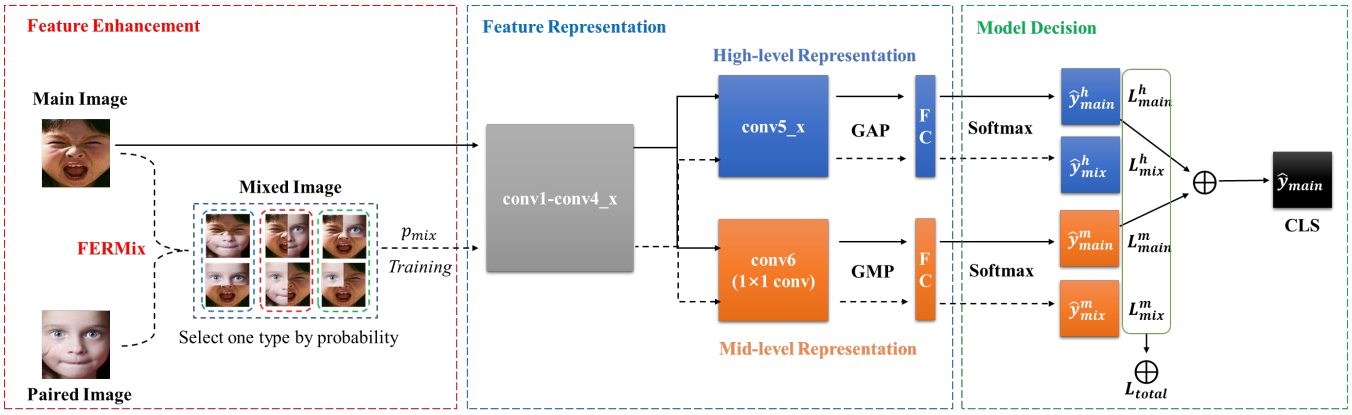


Fig. 2: The framework of the proposed FERMixNet. It divides to three modules, feature enhancement, feature representation, and model prediction. In the **training phase**, the mixed image generated by FERMix is co-trained with the main image. Note that FERMix has a total of six possible sample types, while in practice, only one type is selected as model input for each training iteration by probability. The high-level and mid-level representation branches learn the high-level and mid-level features of the two input samples, respectively. Finally, the final decision result is obtained by adding the high-level and mid-level learning decision results of the main image directly. It is worth mentioning that the mixed image is not involved in the **inference phase**.

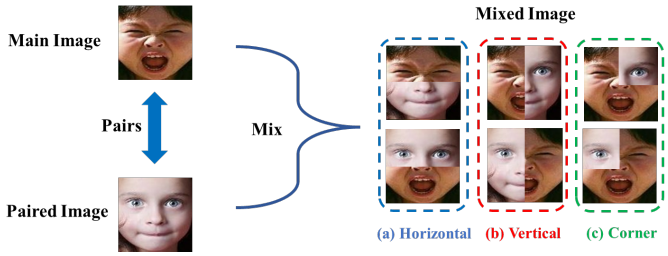


Fig. 3: Overview of the proposed FERMix. There are three patterns of FERMix: (a) horizontal equal mixing; (b) vertical equal mixing; (c) corner mixing. Each pattern contains two types of mixed images, resulting in a total of six possible sample types.

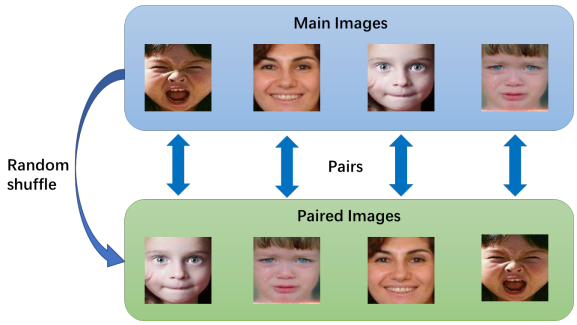


Fig. 4: At each training iteration, the paired images are obtained by random shuffling of the main images. Note: The figure shows an example of batch size is 4.

$\boldsymbol{B}_{b_{left}}$ indicates the mixed region as the left half of the image, representing the left face, while $\boldsymbol{B}_{b_{right}}$ indicates the mixed region as the right half of the image, representing the right face. This mixing pattern simulates the scenario of left and right facial occlusion. Left facial occlusion often results in a large visible region on the right side, while right facial occlusion leads to a large visible region on the left side. By generating samples through vertical equal mixing, the model can effectively focus on the left or right half of the face, thus enhancing the model's robustness against left and right occlusions.

(c) **Corner Mixing**. The mixing pattern corresponds to the left and right eye regions. The left and right eyes tend to be located in the upper left and upper right corners of the face image. Combining the mixing patterns of Fig. 3(a) and Fig. 3(b), our corner mixed region is chosen $\frac{1}{4}$ the upper-left and upper-right regions of the image size. The mixed regions are defined as Eq. (3) shows.

$$\boldsymbol{B}_{c_{left}} = \left(0, 0, \frac{W}{2}, \frac{H}{2}\right), \quad \boldsymbol{B}_{c_{right}} = \left(\frac{W}{2}, 0, \frac{W}{2}, \frac{H}{2}\right) \tag{3}$$

$\boldsymbol{B}_{c_{left}}$ indicates that the mixed region is the upper left part of the image or the left eye. $\boldsymbol{B}_{c_{right}}$ means that the mixed region is the upper right part of the image or the right eye. This type of mixing is intended for cases where only one eye is occluded, such as wearing a single eye patch, hair obscuring, etc.

Let $x_A \in R^{(W\times H\times C)}$ and y_A denote a main image and its label respectively, $x_B \in R^{(W\times H\times C)}$ and y_B denote a paired image and its label respectively. A mixed image $\widetilde{x} \in R^{(W\times H\times C)}$ and its label $\widetilde{y}$ are expressed as Eqs. (4) and (5) show.

$$\widetilde{x} = \boldsymbol{M} \odot x_A + (1 - \boldsymbol{M}) \odot x_B \tag{4}$$

$$\widetilde{y} = \lambda y_A + (1 - \lambda)\, y_B \tag{5}$$

$\boldsymbol{M} \in \{0, 1\}^{W\times H}$ in Eq. (4) denotes a binary mask indicating where to delete and fill in from two images, and $\odot$ is element-wise multiplication. The mixed region $\boldsymbol{B}$ samples the binary mask $\boldsymbol{M}$, i.e., for $\boldsymbol{M}$, the values assigned in region $\boldsymbol{B}$ are 0 and the values in the rest of the region are 1. And the Eq. (4) denotes that the deleted region of the main image is filled with the corresponding region of the paired image. Eq. (5) combines the label information of two original images, with $\lambda \in \{0.5, 0.75\}$. In case of horizontal equal mixing and vertical equal mixing, the main image occupies

half the area of the mixed image and λ takes the value of 0.5. In case of corner mixing, the main image occupies $\frac{3}{4}$ the area of the mixed image and λ takes the value of 0.75. The mixed label combines the label information from two original images, which helps the model better identify the unique features of each image.

Note that each FERMix pattern contains two cases, so there are six types of generated samples in total. While in practice, only one type is selected as the input of each training iteration by probability. We define the probability of using Fig. 3(a) to be 40%, where the mixed region $B_{a_{up}}$ and $B_{a_{down}}$ with the same probability of 20%. The probability of using Fig. 3(b) is 40% where $B_{b_{left}}$ and $B_{b_{left}}$ with the same probability of 20%. And 20% probability of using Fig. 3(c), where $B_{c_{left}}$ and $B_{c_{right}}$ with the same probability of 10%. We illustrate the reason for such a setup in the ablation experiment in Section 4.3.3.

3.2.2 Co-Training the Main Image and the Mixed Image

As shown in Fig. 2, the input of FERMixNet contains two parts, the main image and the mixed image. This paper employs a co-training approach to simultaneously train the main image and the mixed image. This approach is adopted to mitigate the problem of category confusion that may arise when training mixed images separately. On the one hand, the model may combine two parts of the mixed image as the single expression. On the other hand, when using vertical or horizontal equal mixing, each image in the mixed image has half of the weights, which make it difficult for the model to distinguish between the two categories. Co-training addresses the problem by providing the model with supervised information from the main image during training, which helps it differentiate between the two images in the mixed image. Additionally, the local information contained in the mixed image enhances the local feature representation of the main image. Thus, the proposed co-training approach avoids the confusion problem and ensures the FERMix strategy is effective.

In the training phase, we set the input probability of the mixed image as p_{mix}. A random value r is generated from the uniform distribution $(0, 1)$. If $r < p_{mix}$, the model input includes both the main image and the mixed image. If $r \geq p_{mix}$, the model input only contains the main image. We use the dashed line to represent the feedforward process for the mixed image in the Fig. 2. With the experiment in Section 4.3.5, we set $p_{mix} = 0.8$.

It should be noted that the mixed image is not involved in the inference phase.

3.3 Feature Representation

As shown in Fig. 2, the structure of high-level representation learning contains $conv1$, $conv2\_x$, $conv3\_x$, $conv4\_x$, $conv5\_x$, global average pooling (GAP) layer, and fully connected (FC) layer, which are used for learning the global features of the input samples. To enhance the model's capability of learning the non-occluded local features, we add a mid-level representation learning branch after $conv4\_x$, which contains 1×1 convolutional layer, ReLU activation function, global max pooling (GMP) layer, and fully connected (FC) layer. Denote the feature map obtained after $conv4\_x$ as $F \in R^{(W \times H \times C)}$, the equation is written as Eq. (6) shows.

$$F_{mid} = GMP\left(ReLU\left(Conv6\left(F\right)\right)\right) \tag{6}$$

$F_{mid} \in R^{(1 \times 1 \times C)}$ denotes the acquired mid-level features, and Conv6(·) denotes the 1 × 1 convolution operation, and GMP(·) denotes the global max pooling operation. A 1×1 convolutional filter is considered as a detector for a small region. Specifically, the input image generates a $C \times H \times W$ feature map after $conv1 - conv4\_x$, and each $C \times 1 \times 1$ vector across channels reflects the receptive field of a small region at the corresponding location in the original image. In order to locate the discriminative region precisely, the stride of the 1×1 convolutional filter is set to 1. With well-trained 1x1 convolution filters, there is a high degree of responsiveness to important local regions, while global max pooling (GMP) can identify the most discriminative local regions and filter out useless information. In the case of occluded facial expressions, the facial area is important information while the occluded area is irrelevant. By leveraging the mid-level representation learning branch, local features with the highest expression semantic information of the face are obtained and the useless occlusion features are filtered out. Therefore, utilizing the learned mid-level feature representations to complement the high-level feature representations, the impact of occlusion on model performance is minimized, ultimately improving the model's occlusion robustness.

3.4 Model Decision

The losses for both high-level representation learning and mid-level representation learning are calculated using cross-entropy loss, which can be expressed as Eqs. (7) and (8) show:

$$L^h_{main} = -\sum_{i=0}^{C-1} y_i log\hat{y}^h_{main_i}, L^m_{main} = -\sum_{i=0}^{C-1} y_i log\hat{y}^m_{main_i} \tag{7}$$

$$L^h_{mix} = -\sum_{i=0}^{C-1} \widetilde{y}_i log\hat{y}^h_{mix_i}, L^m_{mix} = -\sum_{i=0}^{C-1} \widetilde{y}_i log\hat{y}^m_{mix_i} \tag{8}$$

where the superscript h and m denote the high-level representation learning and the mid-level representation learning respectively. The subscript $main$ and mix denote the main image classification and the mixed image classification respectively. C means the total number of expression classes. y and $\widetilde{y}$ are the ground truth of the main image and the mixed image. And $\hat{y}$ is the prediction result of the model.

During training, the input probability of the mixed image is p_{mix}. Based on it, the total training loss is defined as:

$$L_{total} = \begin{cases} \alpha\left(\beta L^h_{mix} + (1-\beta)\, L^m_{mix}\right) + \\ (1-\alpha)\left(\beta L^h_{main} + (1-\beta)\, L^m_{main}\right), & r < p_{mix} \\ \beta\, L^h_{main} + (1-\beta)\, L^m_{main}, & r \geq p_{mix} \end{cases} \tag{9}$$

where $\alpha \in (0, 1)$ is for controlling the loss weights of the mixed and main image. $\beta \in (0, 1)$ is used to control the loss weights of the high-level and mid-level learning. With the experiments in Section 4.3, we set $\alpha = 0.5$ and $\beta = 0.5$ by default.

For model prediction, we perform decision fusion, i.e., the output units obtained from the main image after high-level and mid-level representation learning branch are

summed up as the final decision result of the model. Note that the mixed image is not involved in model prediction.

4 EXPERIMENTS

To validate the correctness and robustness of the proposed method in realistic facial occlusions, we evaluate it on four public facial occlusion datasets: Occlusion-RAF-DB [13], Occlusion-FERPlus [13], Occlusion-AffectNet [13] and FED-RO [12]. To verify the generality of the proposed model, we also evaluate it on three in-the-wild facial expression datasets: RAF-DB [3], FERPlus [4], and AffectNet [5].

4.1 Datasets

4.1.1 RAF-DB

RAF-DB (Real-world Affective Faces DataBase) [3] contains 29,672 real-world face images with basic or compound expressions. All images were collected from the Internet, and the subjects vary greatly in age, gender, race, lighting, skin color, etc. In this paper, we use 15,339 basic expression images, with 12,271 images used for training and 3,068 for testing. The dataset contains 7 basic expression classes: happy, surprise, sad, anger, disgust, fear, and neutral.

4.1.2 FERPlus

FERPlus [4] is an extension of the standard FER2013 [29]. The dataset contains 35887 grayscale facial expression images of size 48 × 48, with 28709 training images, 3589 validation images, and 3589 test images. There are 8 expression classes, and contempt is added compared with RAF-DB. FERPlus allows 10 annotators to label each image, so each image have 10 annotation information. Following reference [13], we use the maximum voting to decide the ground truth for each image.

4.1.3 AffectNet

AffectNet [5] is one of the largest datasets of facial expressions in the wild, containing 440,000 face images. The dataset contains two benchmarks: AffectNet-7 and AffectNet-8. AffectNet-7 contains 7 basic expression classes, the same as RAF-DB, with a training set of 283,901 and a test set of 3,500. AffectNet-8 contains 8 expression classes, with the additional contempt, of which 287,651 training images and 4,000 test images. Since this dataset suffers from a serious problem of imbalanced samples, we take data down sampling to alleviate this problem.

4.1.4 Occluded Facial Expression Datasets

To validate the performance of expression recognition models in real-world occlusion situations, Li et al. [12] collected and annotated a FED-RO (Facial Expression Dataset with Real Occlusion) in the wild for evaluation. This dataset contains 400 face images of various occlusion situations in real scenes. Following [12], we jointly train the model with the training set of RAF-DB and AffectNet, and evaluate the performance on FED-RO.

Wang et al. [13] built occlusion test datasets Occlusion-RAF-DB, Occlusion-FERPlus, and Occlusion-AffectNet from RAF-DB, FERPlus, and AffectNet, containing 735, 605, and 682 test images, respectively. These subsets are selected from the validation or test datasets of the original datasets with occlusion situation.

In this paper, we divide occlusion into three types: upper and lower occlusion, left and right occlusion, and corner occlusion (as shown in Fig. 1). To illustrate the effectiveness of each pattern of FERMix, we label 735 test images with occlusion types using the Occlusion-RAF-DB dataset. After labeling, the Occlusion-RAF-DB dataset contains 498 upper and lower occlusion images (Occlusion-RAF-DB-UL), 149 left and right occlusion images (Occlusion-RAF-DB-LR), and 88 corner occlusion images (Occlusion-RAF-DB-C).

4.1.5 Pose Variant Facial Expression Datasets

To evaluate the robustness of the face expression recognition model to pose variations, Wang et al. [13] built pose variant test datasets Pose-RAF-DB, Pose-FERPlus, and Pose-AffectNet from RAF-DB, FERPlus, and AffectNet. The pose variations are classified into two types based on the pitch or yaw angles, i.e., larger than 30° and larger than 45°. Among them, Pose-RAF-DB contains 1248 samples with angles larger than 30° and 558 samples with angles larger than 45°, Pose-FERPlus contains 1171 samples with angles larger than 30° and 634 samples with angles larger than 45°, and Pose-AffectNet contains 1949 samples with angles larger than 30° and 985 samples with angles greater than 45°.

4.2 Implementation Details

All of our data is directly adopted from the official face-aligned samples and resized to 224 x 224. We use ResNet18 [26] as the backbone model, which is pre-trained on the Ms-Celeb-1M [30] dataset, as in most studies. It is worth noting that in Section 4.7, to compare with some facial expression recognition models that use a larger backbone, we also use IR50 [31] as the backbone in our experiments.

The RAF-DB and FERPlus datasets are trained for 50 epochs, and augmented by random horizontal flipping and random cropping. The batch size is set to 64. The model parameters are optimized via momentum stochastic gradient descent (SGD) optimizer, with initial learning rate of 0.1, momentum of 0.9 and weight decay of 10^{-4}. The learning rate is decreased by 10 every 10 epochs.

The AffectNet dataset is also trained with 50 epochs. The data is augmented with random horizontal flipping and random affine transformation, where the scale value ranges from 0.8 to 1.0 and the translate value is 0.2. To cope with the problem of imbalanced samples, we use down sampling, i.e., we reduce the training samples from the classes with relatively more samples in the training set. The batch size is set to 16. The model parameters are optimized via Adam optimizer with weight decay of 10^{-4}. The initial learning rate is set to 0.0001 and exponentially decayed by a factor of 0.8 every epoch.

For some hyperparameters in our method, the input probability of the mixed image p_{mix} is set to 0.8. In Eq. (9), α and β are both set to 0.5. We will elaborate on this later in the ablation experiments.

We implement the proposed model using the Pytorch framework, and all experiments are done with a NVIDIA GTX 3060Ti GPU.

4.3 Ablation Studies

In this section we demonstrate the validity of the FERMix and mid-level representation learning through a series of ablation studies, as well as illustrate the settings of the individual hyperparameters.

4.3.1 Validity of FERMix and Mid-Level Representation Learning

TABLE 1
Ablation results for FERMix and mid-level representation learning on RAF-DB
(The comparison metric is the number of paramters, GFLOPs and accuracy. ↑ indicates that the metric is better as it gets higher, the value in bold indicates the best result)

FERMix	Mid-level	#Params(M)	#GFLOPs	Accuracy (%) ↑ Occ-RAF-DB	RAF-DB(Full)
-	-	11.18	1.82	82.99	88.04
✓	-	11.18	1.82	85.99	89.24
-	✓	11.25	1.83	85.99	89.02
✓	✓	11.25	1.83	**86.67**	**89.86**

In this subsection, we conduct experiments on the RAF-DB dataset to evaluate the impact of FERMix and mid-level representation learning on the model performance, as summarized in TABLE 1. When FERMix and mid-level representation learning are not employed, the accuracy achieved on the occlusion test set and the full test set is 82.99% and 88.04%, respectively. However, when the mixed images generated by FERMix are co-trained with the main images without the use of mid-level representation learning, the accuracy improves to 85.99% (+3.00%) on the occlusion test set and 89.24% (+1.20%) on the full test set. These results suggest that the FERMix strategy improves the occlusion robustness of the model and enhances its overall generalization. As FERMix is a data augmentation strategy, it does not increase the number of parameters or GFLOPs.

Here we explain why the FERMix strategy is effective. During model training, the model's ability to identify expressions based on complete facial information is stronger, owing to the majority of non-occluded face samples in the training dataset. Conversely, the model's ability to recognize expressions based on non-complete facial information is weaker, as the training data lacks non-complete face samples. Consequently, when an occluded face is encountered in the test sample, leading to an incomplete face, the model's performance in recognizing expressions based on incomplete face information becomes weaker, resulting in poor performance in handling occlusion situations. However, FERMix generates mixed samples by augmenting incomplete face samples to the training set, thereby improving the model's ability to recognize expressions using incomplete facial information. As a result, the FERMix strategy effectively enhance the occlusion robustness of the model. Additionally, since the FERMix strategy increases the diversity of input samples, it further enhances the generalization of the model.

In the absence of the FERMix strategy, incorporating mid-level representation learning improves the model's accuracy on the occlusion test set to 85.99% (+3.00%) and on the full test set to 89.02% (+0.98%). This is because the discriminative features learned in the mid-level representation learning branch effectively highlight important regions and suppress the impact of occlusion information, thereby mitigating the negative impact of occlusion on the model's performance. The model only has a slight increase in the number of parameters (+0.63%) and GFLOPs (+0.55%). These results indicate that mid-level representation learning enhances the model's occlusion robustness and generalization at a low cost.

When employing both the FERMix strategy and mid-level representation learning, the accuracy on the occlusion test set is further enhanced to 86.67% (+3.68%), and 89.86% (+1.82%) on the full test set. The results suggest that there is no conflict between the two methods, and their combined use can lead to further improvement in the occlusion robustness and generalization of the model.

4.3.2 Validity of Each FERMix Pattern for Different Occlusion Types

TABLE 2
Ablation results for the three FERMix patterns on Occlusion-RAF-DB-UL, Occlusion-RAF-DB-LR, Occlusion-RAF-DB-C
(The comparison metric is accuracy. ↑ indicates that the metric is better as it gets higher, the value in bold indicates the best result. The experimental model is vanilla ResNet18 without mid-level representation learning.)

Method	Accuracy (%) ↑ Occ-RAF-DB-UL	Occ-RAF-DB-LR	Occ-RAF-DB-C
Baseline	84.54	81.88	78.41
Horizontal(a)	**85.34**	82.55	80.68
Vertical(b)	84.94	**83.22**	82.59
Corner(c)	84.74	79.87	**86.36**

To assess the effectiveness of each FERMix pattern for different occlusion types, we divide the Occlusion-RAF-DB test set according to the occlusion type and conduct experiments on each subset. The subsets are named as Occlusion-RAF-DB-UL, Occlusion-RAF-DB-LR, and Occlusion-RAF-DB-C, representing upper and lower, left and right, and corner occlusions, respectively.

TABLE 2 displays the outcomes of our experiments, where baseline denotes the performance achieved without using any FERMix pattern. We find that training with the horizontal equal mixing pattern yielded the best results on the upper and lower occlusion data, resulting in a performance gain of 0.80% compared to the baseline. Furthermore, training with the vertical equal mixing pattern achieve the best accuracy on the left and right occlusion data, displaying a 1.34% improvement over the baseline. The corner mixing pattern demonstrate the highest accuracy on the corner occlusion data, resulting in a substantial improvement of 7.95% compared to the baseline. Our findings indicate that using the three FERMix patterns improve the model's robustness against the three types of occlusions by simulating real-world occlusion scenarios.

4.3.3 Sampling Probabilities of Three mixing Patterns

In order to determine the optimal sampling probabilities of three mixing patterns, we set different sampling probabilities for three mixing patterns to explore their effects on model occlusion robustness and generalization. Due to the numerous possible probability scenarios for the three mixing methods, in order to obtain systematic results within a limited number of groups, we use corner mixing as the baseline and set the probabilities at 40%, 30%, 20%, and 10% for four major groups. And each group containing three cases: horizontal mixing probability greater than vertical, horizontal mixing probability less than vertical, and horizontal mixing probability equal to vertical. Thus, there are 12 mixing probability combinations. We use the RAF-DB dataset, and train a vanilla ResNet18 model without mid-level representation learning. The results are presented in the TABLE 3.

TABLE 3
Ablation results for different FERMix sampling probabilities on Occlusion-RAF-DB and RAF-DB(Full) (↑ indicates that the metric is better as it gets higher, the value in bold indicates the best result. The experimental model is vanilla ResNet18 without mid-level representation learning.)

Probability (%)			Accuracy (%) ↑	
Horizontal	Vertical	Corner	Occ-RAF-DB	RAF-DB (Full)
40	20	40	84.90	88.95
20	40	40	85.03	88.66
30	30	40	85.17	89.05
40	30	30	85.31	89.15
30	40	30	85.17	89.08
35	35	30	85.44	89.21
60	20	20	85.71	89.08
20	60	20	85.58	89.08
40	40	20	**85.99**	**89.24**
80	10	10	85.03	88.66
10	80	10	84.90	88.72
45	45	10	85.31	89.05

From the analysis of the four major groups, it is observed that as the corner mixing probability decreases, the model's occlusion robustness and generalization tend to improve, reaching an optimal point at 20%. This indicates that the sampling probability for corner mixing should be lower than that for horizontal and vertical mixing, ideally around 20%. Within each group, it is evident that the model achieves better occlusion robustness and generalization when the horizontal and vertical mixing probabilities are equal. Based on these findings, we recommend setting the probabilities of horizontal, vertical, and corner blending to 40%, 40%, and 20%, respectively.

4.3.4 Different Combinations of Three FERMix Patterns

In this subsection, we conduct ablation experiments on different combinations of three FERMix patterns to demonstrate their effectiveness. We use the RAF-DB dataset, and train a vanilla ResNet18 model without mid-level representation learning. The results of the ablation experiments are shown in TABLE 4.

TABLE 4
Ablation results for the three FERMix patterns on Occlusion-RAF-DB and RAF-DB(Full) (The comparison metric is accuracy. ↑ indicates that the metric is better as it gets higher, the value in bold indicates the best result. The experimental model is vanilla ResNet18 without mid-level representation learning.)

Horizontal	Vertical	Corner	Accuracy (%) ↑	
			Occ-RAF-DB	RAF-DB(Full)
-	-	-	82.99	88.04
✓	-	-	84.22	88.30
-	✓	-	84.35	88.59
-	-	✓	83.95	88.56
✓	✓	-	85.85	89.02
✓	-	✓	85.71	88.89
-	✓	✓	85.58	88.69
✓	✓	✓	**85.99**	**89.24**

Rows 1-4 of the table demonstrate that each pattern of FERMix improves the occlusion robustness and generalization of the model, as the accuracy of the model improves on both the occlusion test set and the full test set when each pattern is used alone.

Rows 5-7 of the table show the results of the two-by-two combination of FERMix patterns. The experimental results indicate that the addition of one more FERMix pattern further improve the performance of the model, as it increases the richness of the training samples.

The last row of the table shows the results of combining all three FERMix patterns. Based on the findings from Section 4.3.3, we set the probabilities of horizontal, vertical, and corner mixing to 40%, 40%, and 20% respectively. The experimental results demonstrate that combining all three patterns further increases the richness of the samples and improves the model performance. Besides the three existing patterns, we may also consider additional patterns, such as diagonal mixing. However, we believe that our current set of three patterns sufficiently cover the occlusion issues encountered in real-world scenarios. Therefore, we have not integrated more mixing patterns.

4.3.5 The Role of Co-training

In Section 3.2.2, co-training is introduced where both the main image and the mixed image are simultaneously input to the network for training in one iteration. Another method is to train using only the mixed or main image in one iteration depending on the probability. This subsection presents experimental results for these two training methods to illustrate the necessity of co-training. We use the RAF-DB dataset, and preset the weight parameters $\alpha = 0.5$ and $\beta = 0.5$ in Eq. (9).

In Fig. 5, the blue and orange solid lines represent the results obtained using co-training and without co-training under different probabilities of mixed image inputs. It can be seen that the accuracy of using co-training is higher under any probability of mixed image inputs. And the orange solid line exhibits a decline in model performance when co-training is not utilized, except for the cases where p_{mix} is 0.4 and 0.5. This is because when the FERMix pattern is vertical or horizontal equal mixing, both images occupy equal proportions of the mixed image, and the labels

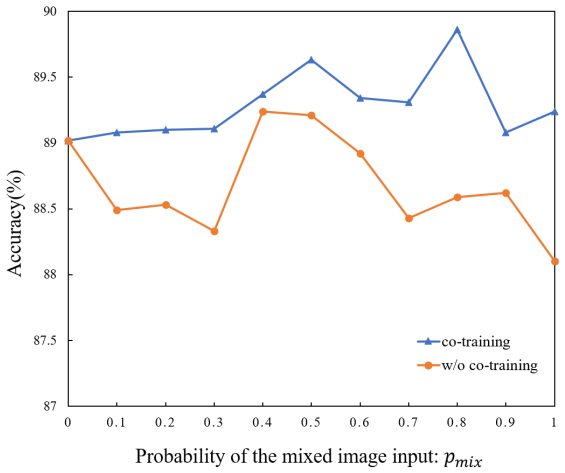


Fig. 5: Evaluation of different p_{mix} value and training strategy

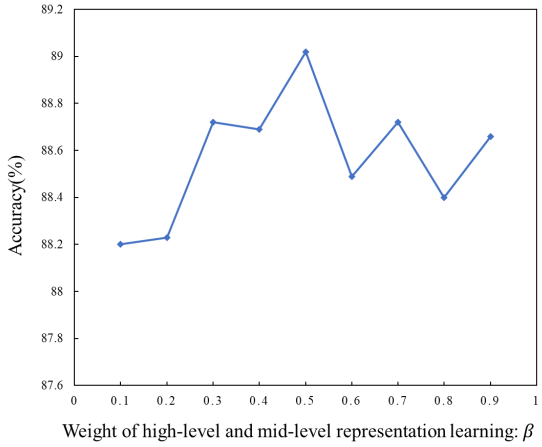


Fig. 6: Evaluation of different β values for high-level and mid-level representation learning

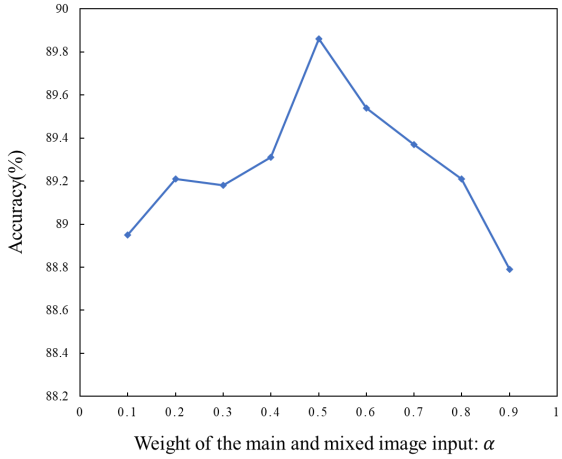


Fig. 7: Evaluation of different α values for the main and mixed image

of the mixed image are also equally distributed between the categories of the main and paired image. Therefore, the model is likely to confuse the categories of the two images. To address this issue, we propose a co-training approach. During training, both the main and mixed images are input together, enabling the model to receive supervised information from the main image and better distinguish between the two images in the mixed image. Meanwhile, this method enhances the local feature representation of the main image by incorporating the local information from the mixed image. Upon closer examination of the blue solid line, it is evident that after implementing co-training, the model's performance improves regardless of the input probability p_{mix} of the mixed image, and the optimal performance of 89.86% is achieved when p_{mix} is set to 0.8. This indicates that the co-training technique effectively leverages the FERMix strategy.

In summary, we set p_{mix} to 0.8 and utilize co-training to avoid confusion problem of the model, allowing the FERMix strategy to be effectively utilized.

4.3.6 *Weighting Parameter β for High-Level and Mid-Level Representation Learning*

This subsection explores the significance of the weighting parameter β, which controls the balance between high-level and mid-level representation learning. To eliminate the impact of α, we set p_{mix} to 0 and do not feed the mixed image to the model. Fig. 6 illustrates the accuracy trend as β varies from 0.1 to 0.9. According to Eq. (9), a larger β emphasizes high-level representation learning. When β is 0.1, the model ignores high-level representation and has the lowest accuracy of 88.20%. This indicates that the high-level representation is essential in facial expression recognition. When $\beta = 0.5$, the model performance is the best, as both high-level representation and mid-level representation are learned well, and mid-level representation provides a better complement to high-level representation. Therefore, we set β to 0.5.

4.3.7 *Weighting Parameter α for the Main and Mixed image*

This subsection investigates the impact of the weighting parameter α on the model's performance, which controls the training of the main image and the mixed image. After the previous ablation experiments, with $\beta = 0.5$ and $p_{mix} = 0.8$ fixed, we examine the accuracy rate while varying α from 0.1 to 0.9. According to Eq. (9), a higher value of α indicates a greater emphasis on training the mixed image. According to Fig. 7, the model's performance is the worst when $\alpha = 0.9$, with an accuracy of only 88.79%. At this point the model focuses much more on the training of the mixed image and ignores the training of the main image. This suggests that ignoring the global information affects the model performance. Conversely, when $\alpha = 0.5$, both the main image and mixed image are trained well, and the model achieves the highest accuracy of 89.86%. We think that training the main image helps the mixed image in utilizing local information, while the mixed image, in turn, enhances the model's ability to represent local information in the main image. As a result, we set α to 0.5 as it strikes a balance between the two types of images.

4.4 Comparison with Some Cutting-based and Mixing-based Data Augmentation Strategies

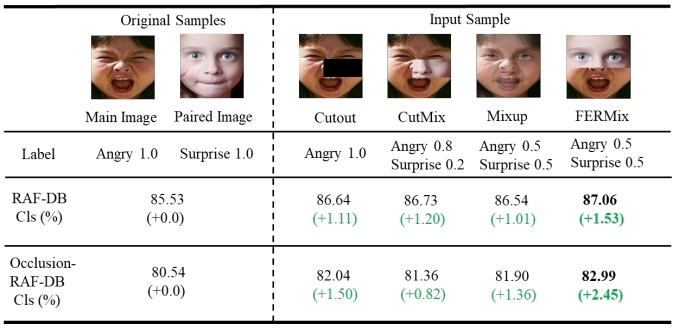


Fig. 8: Comparison of Cutout, CutMix, Mixup, and FERMix. Baseline result is shown on the left side of the dashed line. And the results of three data augmentation are shown on the right side of the dashed line. CutMix mixes the nose part region of the paired image with the main image, but the nose part region contains only little information about the expression, so such mixed image may affect the performance of the model.

In order to demonstrate the high adaptability of FERMix for the facial expression dataset, we have conducted a comparison study with several common cutting-based and mixing-based data augmentation strategies, including Cutout, CutMix, and Mixup. To more accurately evaluate the effects of these strategies on the experimental results, we do not apply additional data augmentation operations, such as random flipping or cropping, to the dataset. In this section, we employ a vanilla ResNet18 model and utilize co-training approach for each data augmentation strategy to ensure a fair comparison. The experiments are conducted on the RAF-DB dataset, and we compare the classification accuracies of the four strategies on both the full and occlusion test sets. The results are presented in Fig. 8.

Results show that all four strategies are able to improve the performance of the facial expression recognition model, indicating that image cutting or mixing augmentation strategies can improve the generalization and occlusion robustness of the model. Among these methods, FERMix achieves the best results.

Cutout randomly changes a region of pixels of the image to zero to simulate the occlusion situation, but this may result in the cutting region becoming useless information and reduce the information richness of the training samples, thereby decreasing the training efficiency of the model. FERMix replaces the cutting region with pixel values from the same position region of another face image, which preserves the information of the training sample and enhances the efficiency of the model training. Additionally, since the mixed image contains information from different images, the model can better learn the similarities and differences between different images, thereby improving the generalization ability of the model.

CutMix randomly selects regions for image replacement, which may result in low semantic information richness of the selected regions. This can affect the performance of the model. In contrast, FERMix is tailored to the characteristics of facial expression data and selects local regions with higher expression semantic information for mixing. This can improve the model's performance on the facial expression dataset.

Mixup is a linear interpolation of two images and labels to generate a new sample and a new sample label to enhance the robustness of the model to noise. Since occluded data can be treated as a special case of noise data, Mixup can also improve the robustness of the model to occlusion. However, the mixed images generated by Mixup may not be representative of real-world face occlusion cases. FERMix simulates three common face occlusion cases in the real world and further improves the model's occlusion robustness.

4.5 Comparison on the Occluded Facial Expression Datasets

FERMixNet is aimed to improve the occlusion robustness of facial expression recognition. To verify the effectiveness of FERMixNet, we evaluate the proposed model on four occlusion facial expression datasets, FED-RO, Occlusion-RAF-DB, Occlusion-FERPlus and Occlusion-AffectNet. We compare the FERMixNet with several state-of-the-art methods for occlusion facial expression recognition, including gACNN [12], RAN [13], VTFF [14], EfficientFace [15], MAPNet [10], MViT [9], LAENet-SA [16], and Co-Completion [11]. In addition to comparing the accuracy of the models, we also evaluated two complexity metrics: the number of parameters and GFLOPs. However, since some methods do not report these metrics, and their implementation is not publicly available, we only provide lower-bound estimates (the symbol $>$ is used) based on their model structure.

TABLE 5
Comparison on FED-RO
(The comparison metric is the number of paramters, GFLOPs and accuracy. * denotes the method using VGG-16 as backbone, † denotes the method using ShuffleNet-V2 as backbone, otherwise the backbone is ResNet18. ↑ indicates that the metric is better as it gets higher, ↓ indicates that the metric is better as it gets lower, the value in bold indicates the best result, the underlined indicates the second-best result.)

Method	#Params(M) ↓	#GFLOPs ↓	Accuracy(%) ↑
gACNN* [12]	>134.29	>15.48	66.50
RAN [13]	11.19	14.55	67.98
EfficientFace† [15]	**1.28**	**0.15**	68.25
LAENet-SA [16]	11.20	2.01	68.25
MAPNet [10]	>11.25	>1.83	71.50
Co-Completion [11]	>25.56	>3.53	<u>72.50</u>
FERMixNet(Ours)	11.25	<u>1.83</u>	**72.97**

TABLE 5 presents the comparison results on the FER-RO. We do not list the results of VTFF and MViT since it is not evaluated on this dataset and the implementation is not publicly available. The proposed model achieves the best accuracy on FER-RO. This demonstrates the effectiveness of the mixed samples generated by FERMix to simulate occlusion and the local discriminative features learned by the mid-level representation learning, which enhances the model's occlusion robustness. When comparing model

TABLE 6
Comparison on Occlusion-RAF-DB, Occlusion-FERPlus, and Occlusion-AffectNet
(The comparison metric is the number of paramters, GFLOPs and accuracy. ↑ indicates that the metric is better as it gets higher, ↓ indicates that the metric is better as it gets lower, the value in bold indicates the best result, the underlined indicates the second-best result)

Method	#Params(M) ↓	#GFLOPs ↓	Accuracy (%) ↑		
			Occ-RAF-DB	Occ-FERPlus	Occ-AffectNet
RAN [13]	<u>11.19</u>	14.55	82.72	83.63	58.50
EfficientFace [15]	**1.28**	**0.15**	83.24	-	59.88
VTFF [14]	51.80	>4.60	83.95	<u>84.79</u>	**62.98**
MViT [9]	>21.80	>4.60	<u>85.17</u>	-	-
FERMixNet(Ours)	11.25	<u>1.83</u>	**86.67**	**85.95**	<u>62.66</u>

complexity, EfficientFace stands out with its lightweight ShuffleNet-V2 backbone [32], which has the lowest model parameters and GFLOPs among all compared methods. However, its performance is limited due to its lightweight architecture. Compared with other methods except EfficientFace, the proposed model has only a slightly larger number of parameters compared to RAN and LAENet-SA, with an increase of 0.06M (+0.54%) and 0.05M (+0.45%) respectively. Howerver, in terms of GFLOPs, the proposed model exhibits a significantly lower computational complexity, with a reduction of 12.72 GFLOPs (-87.42%) compared to RAN and 0.18 GFLOPs (-8.96%) compared to LAENet-SA. In fact, the proposed model ranks second only to EfficientFace in terms of GFLOPs, making it one of the most efficient models. These results indicate that the proposed model exhibits a good tradeoff between model performance and complexity.

TABLE 6 shows the results on Occlusion-RAF-DB, Occlusion-FERPlus, and Occlusion-AffectNet. We do not list the results of gACNN, LAENet-SA, MAPNet and Co-Completion on these datasets, since they do not have experimental results for these datasets and the implementation is not publicly available. The results indicate that FERMixNet outperforms the other methods, achieving the best accuracy of 86.67% and 85.95% on Occlusion-RAF-DB and Occlusion-FERPlus, respectively. Although the performance of the proposed model ranks second on Occlusion-AffectNet, slightly inferior to that of VTFF, FERMixNet has significantly fewer parameters and GFLOPs than that of VTFF, which uses the Vision Transformer module and thus greatly increases the model's complexity.

Among the methods we compare, gACNN, RAN, VTFF, EfficientFace, and LAENet-SA mainly use attention mechanisms to focus on non-occluded local regions in the input samples, while MViT, MAPNet, and Co-Completion mask off the occluded regions to reduce their impact on the model. In contrast, FERMixNet addresses the problem of occlusion mainly through data augmentation, by increasing the number of simulated occlusion samples to improve the model's occlusion robustness. Additionally, data augmentation is only applied during the training phase and does not increase the model's inference time. Therefore, our method can be effectively applied in real-world scenarios. Furthermore, we investigate the use of mid-level representation learning to address the occlusion problem. It is utilized to focus the model on the non-occluded regions in the input image, similar to the attention mechanism does. However, when comparing GFLOPs, the proposed model has lower computational complexity compared with attention-based methods, with the exception of EfficientFace. Notably, EfficientFace uses a lightweight network, ShuffleNet-V2, as its backbone, with a focus on reducing model complexity. This shows that mid-level representation learning is more efficient than attention-based methods.

In summary, the FERMixNet strikes a balance between model performance and complexity, demonstrating that it is a practical and efficient solution for occluded facial expression recognition.

4.6 Comparison on the Pose Variant Facial Expression Datasets

Although the proposed model is designed for occluded facial expression recognition, we also conduct experiments to evaluate its effectiveness on pose variant scenarios. The results show that the proposed model achieve the best performance on Pose-RAF-DB (30), Pose-FERPlus (30), and Pose-AffectNet (45) datasets, and the second-best performance on other test sets. This indicates that the proposed model is also effective to face pose variation. Our analysis suggests that face pose variation usually result in some regions of the face to be invisible, leading to a need for enhanced the visible local feature representation in order to improve model robustness for pose variation. The proposed FERMix strategy increases the number of training samples containing local facial region information, and mid-level representation learning enhance the local feature representation. Thus, the proposed model achieves good results in the presence of pose variations.

4.7 Comparison on In-the-Wild Facial Expression Datasets

To examine the generality of FERMixNet in realistic and complex scenarios, we compare the performance of it with that of the existing state-of-the-art methods on three large in-the-wild facial expression datasets RAF-DB, FERPlus and AffectNet. In addition to some methods [9], [10], [13], [14], [16] mentioned in Section 4.5, SCN [33], DACL [34], RUL [35], DMUE [36], FDRL [37], TransFER [38], Meta-Face2Exp [39], LDLVA [40], EAC [41] are also included.

SCN, DMUE, RUL, EAC, and LDLVA primarily address the label ambiguity issue in facial expression recognition,

TABLE 7
Comparison of accuracy on Pose-RAF-DB, Pose-FERPlus, and Pose-AffectNet
(The comparison metric is accuracy. ↑ indicates that the metric is better as it gets higher, the value in bold indicates the best result, the underlined indicates the second-best result)

Method	RAF-DB (%) ↑		FERPlus (%) ↑		AffectNet (%) ↑	
	Pose (30)	Pose (45)	Pose (30)	Pose(45)	Pose (30)	Pose (45)
ResNet18 [13]	84.04	83.15	78.11	75.50	50.10	48.50
RAN [13]	86.74	85.20	82.23	80.40	53.90	53.19
EfficientFace [15]	<u>88.13</u>	86.92	-	-	57.36	56.87
VTFF [14]	87.97	**88.35**	<u>88.29</u>	**87.20**	**60.61**	<u>61.00</u>
FERMixNet(Ours)	**89.01**	<u>87.63</u>	**88.55**	<u>86.73</u>	<u>60.49</u>	**61.14**

while DACL and FDRL target the problem of high intra-class distance and low class spacing of expressions. Meta-Face2Exp focuses on class imbalance problems in facial expression datasets. And TransFER leverage the Transformer encoder module to learn the relationship of each local region of expressions, leading to significant performance gains.

Experimental results are divided into two groups based on backbone usage: one group uses ResNet18 as the backbone, as shown in TABLE 8(a), while the other group uses a larger model, as shown in TABLE 8(b). For the experimental comparison group of larger models, FERMixNet employs IR50 as the backbone, and training optimization utilizes sharpness-aware minimization to improve model generalization by minimizing both the loss value and loss sharpness [42].

When ResNet18 is used as the backbone, the proposed model achieves competitive results compared to state-of-the-art methods. Specifically, on the AffectNet-7 dataset, the proposed model achieves the best performance. On the RAF-DB and FERPlus datasets, the proposed model's accuracy is slightly lower than that of EAC by 0.13% and 0.06%, respectively, but higher than EAC's on the AffectNet-7 dataset by 0.22%. On the AffectNet-8 dataset, the accuracy of the proposed model is slightly lower than that of DMUE, while on the RAF-DB and FERPlus datasets, it is much better than that of DMUE by 1.04% and 0.94%, respectively. Moreover, when using IR50 as the backbone, the proposed model achieves state-of-the-art results on the RAF-DB, AffectNet-7, and AffectNet-8 datasets, while performing slightly lower than TransFER does under FERPlus. The experimental results demonstrate that by increasing the diversity of training samples through the FERMix strategies, and leveraging the most discriminative local features learned through mid-level representation learning, FERMixNet exhibits strong generalization ability on the in-the-wild facial expression recognition dataset.

Compared with most other state-of-the-art methods that enhance facial expression recognition performance by modifying the model structure, the FERMix strategy improves the model's generalization ability through data augmentation. Thus, the proposed model improve generalization without significantly increasing the model's parameters or computational complexity. Specifically, FERMixNet simply adds a mid-level representation learning branch to the vanilla ResNet and uses the FERMix strategy for training, resulting in a negligible increase in parameters and GFLOPs relative to the vanilla ResNet, as demonstrated in TABLE 1. This simplicity and efficiency make the proposed model highly practical. Furthermore, the FERMix strategy is highly flexible and can be applied as a data augmentation strategy to various facial expression recognition models, enhancing their robustness to occlusion.

4.8 Visualization Analysis

4.8.1 What Does Model Learn with FERMix

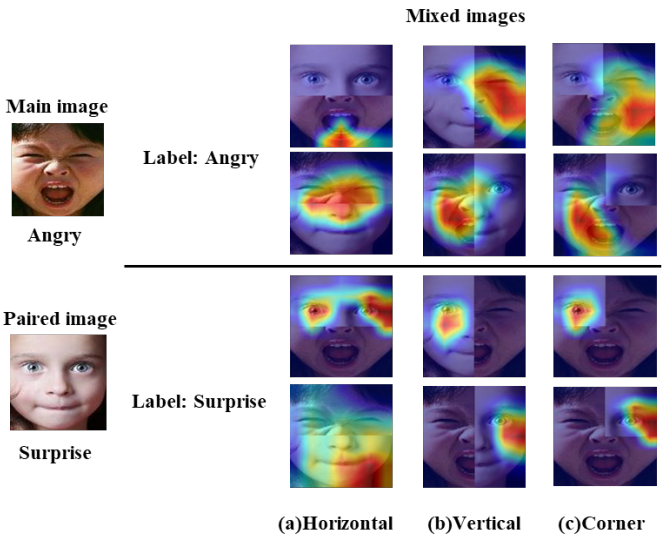


Fig. 9: Class activation mapping (CAM) visualizations on "Angry" and "Surprise" mixed samples. The top of the line shows the visualization of the model when the label is "Angry". And the bottom of the line shows the visualization of the model when the label is "Surprise".

To verify the effectiveness of the FERMix strategy in enhancing the utilization of local information from mixed images, we leverage Grad-CAM [43] to visualize the proposed model. Specifically, we select a mixed image consisting of a main image labeled as "Anger" and a paired image labeled as "Surprise", and examine the model's ability to distinguish these two expressions based on the visualization results. As shown in Fig. 9, the model is able to focus on the main image region when the label is set to "Angry", and on the paired image region when the label is set to "Surprise". These results demonstrate that the proposed model effectively utilize local features from mixed images to distinguish different facial expressions, thus validating

TABLE 8
Comparison of accuracy on RAF-DB, FERPlus, AffectNet-7 and AffectNet-8
(‡ denotes the method using ResNet50 or variant as backbone, $^{+}$ denotes the method using ViT as backbone, otherwise the backbone is ResNet18. ↑ indicates that the metric is better as it gets higher, the value in bold indicates the best result, the underlined indicates the second-best result)

(a) Compare the Methods with ResNet18

Method	Accuracy (%) on Different Dataset↑			
	RAF-DB	FERPlus	AffectNet-8	AffectNet-7
RAN [13]	85.07	-	58.78	-
SCN [33]	86.90	88.55	59.50	-
VTFF [14]	88.14	88.81	61.85	64.80
DACL [34]	87.78	-	-	65.20
RUL [35]	88.98	-	-	-
FDRL [37]	89.47	-	-	-
DMUE [36]	88.76	88.64	**62.84**	-
MAPNet [10]	87.26	-	-	64.09
LAENet-SA [16]	-	-	61.22	64.09
EAC [41]	**89.99**	**89.64**	-	<u>65.32</u>
FERMixNet(ours)	<u>89.86</u>	<u>89.58</u>	<u>61.92</u>	**65.54**

(b) Compare the Methods with a Larger Backbone

Method	Accuracy (%) on Different Dataset↑			
	RAF-DB	FERPlus	AffectNet-8	AffectNet-7
SCN‡ [33]	-	89.35	-	-
DMUE‡ [36]	89.42	89.51	<u>63.11</u>	-
MViT^{+} [9]	88.62	89.22	-	64.57
TransFER‡ [38]	<u>90.91</u>	**90.83**	-	<u>66.23</u>
Meta-Face2Exp‡ [39]	89.47	-	-	64.23
LDLVA‡ [40]	90.51	-	-	66.23
FERMixNet‡(ours)	**91.62**	<u>90.25</u>	**63.22**	**66.40**

the effectiveness of the FERMix strategy in enhancing the utilization of local features.

4.8.2 The Role of Mid-Level Representations for Occluded Facial Expression Recognition

To validate that the mid-level representation learns discriminative local features in non-occluded facial regions, we evaluate the mid-level representation on occluded samples. We visualize the learned representations using Grad-CAM [43], as shown in Fig. 10. Our findings demonstrate that high-level representation learning focuses on global facial information, while mid-level representation learning emphasizes discriminative local regions. We hypothesize that the most discriminative local regions are typically not occluded, and therefore, the mid-level representation learning branch can mitigate the negative effects of occlusion and improve the model's occlusion robustness.

4.8.3 The Role of Co-Training for Occluded Facial Expression Recognition

To illustrate the effect of co-training. As shown in Fig. 11, we take an example of an angry sample and a surprised sample, which are horizontally mixed to produce a mixed image. This mixed image overall looks more like a surprised expression. We used Grad-CAM [43] to analyze the regions of focus under different labels with and without co-training. Without co-training, when the label is surprised, the model considers both local regions of the original image as showing a surprised expression. This is clearly not meeting our expectations, as the whole image does tend to favour the surprised expression. However, with the co-training approach, the model can distinguish the local regions corresponding to the surprised expression because we input the original images during the training process. The model, after learning from the original images, has the ability to differentiate which part of the mixed image belongs to which original image.

4.8.4 Limitations of the proposed method

For the proposed method FERMix, the current implementation is suitable for most aligned facial data. However, it presents limitations when dealing with significant facial occluded data. Occlusions may introduce noise in the mixing areas. As illustrated in Fig. 12, noise in horizontal, vertical, or corner mixing complicates recognition. This issue can be addressed through additional methods such as leveraging facial landmarks to accurately locate critical areas for mixing, albeit at increased training costs. Therefore, it is essential to balance performance and cost to select the most appropriate implementation approach.

5 CONCLUSION

We propose FERMixNet, an occlusion-robust facial expression recognition model that utilizes facial mixing augmentation and mid-level representation learning. To address the challenge of occluded faces, we generate mixed training samples by performing horizontal, vertical, and

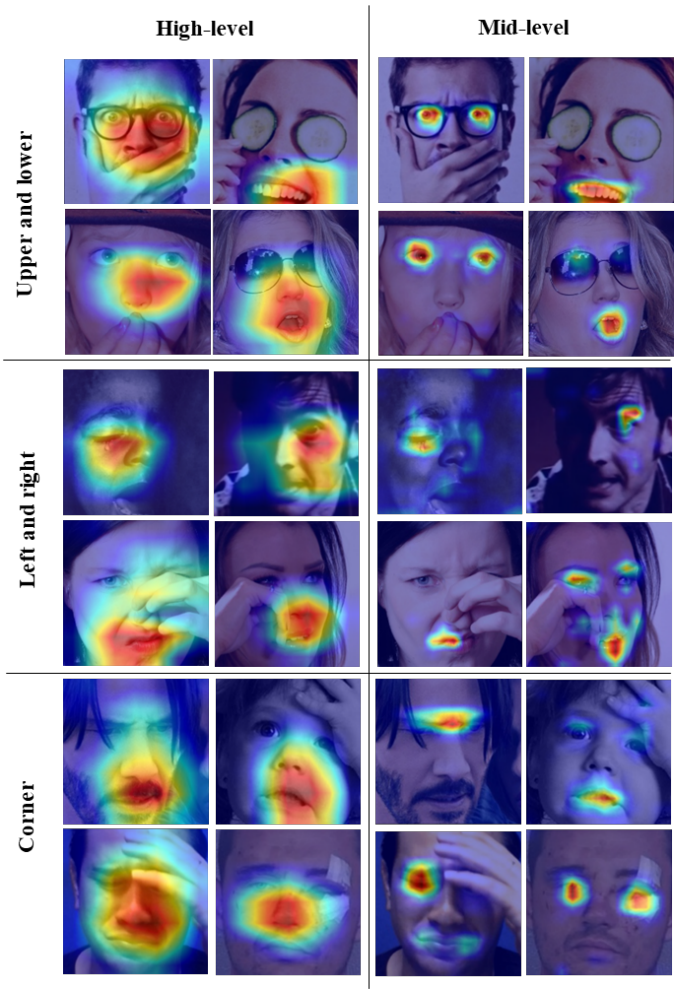


Fig. 10: High-level and mid-level representations learned by the model.

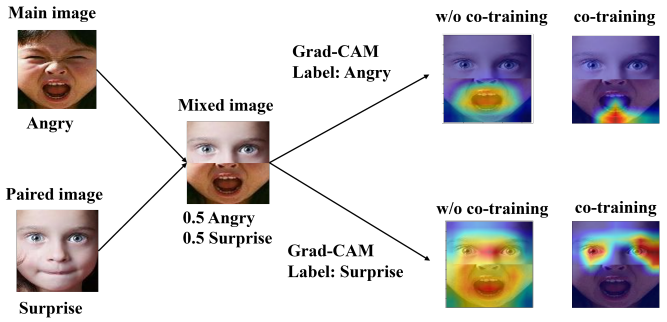
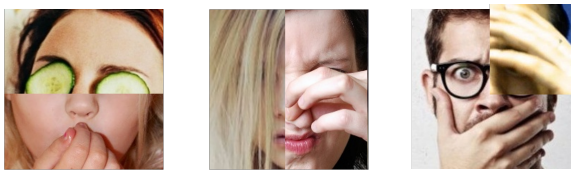


Fig. 11: The effect of co-training.



(a) Horizontal (b) Vertical (c) Corner

Fig. 12: FERMix Failure Cases.

corner mixing on face images, simulating realistic occlusion while preserving high expression-related semantic information. During training, we co-train the main and mixed images to enhance feature representation of local information. Additionally, we introduce a mid-level representation learning branch to capture the most discriminative non-occluded local features of face samples. Extensive experiments are carried out on public datasets. The experimental results demonstrate that the proposed model is effective with high occlusion robustness. It achieves state-of-the-art results on the in-the-wild facial expression datasets RAF-DB, AffectNet-8, and AffectNet-7, outperforming other methods. These results suggest that the proposed model has promising application prospects in the real world.

In future work, we can further optimize the implementation of FERMix to tackle challenges posed by significant variations in facial occlusions. The current implementation is generally suitable for aligned facial data, but may lead to noise problems in the facial mixing region when facing facial occlusions. To address these challenges, additional auxiliary methods can be considered. For instance, leveraging facial landmark information to precisely localize critical regions could improve the accuracy of region mixing in complex scenarios. However, it may also increase training costs and computational complexity, necessitating a balance between performance and cost.

Acknowledgments

This work is supported by the 2022 Shanghai Service Industry Development Fund (No.06162021592), Shanghai Technical Service Center of Science and Engineering Computing, Shanghai University.

References

[1] C. Darwin and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[2] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[3] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2852–2861.

[4] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 279–283.

[5] 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.

[6] Y. Lu, S. Wang, W. Zhao, and Y. Zhao, "Wgan-based robust occluded facial expression recognition," *IEEE Access*, vol. 7, pp. 93 594–93 610, 2019.

[7] M. Ranzato, J. Susskind, V. Mnih, and G. Hinton, "On deep generative models with applications to recognition," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 2857–2864.

[8] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[9] H. Li, M. Sui, F. Zhao, Z. Zha, and F. Wu, "Mvit: Mask vision transformer for facial expression recognition in the wild," *CoRR*, vol. abs/2106.04520, 2021.

[10] L. Ju and X. Zhao, "Mask-based attention parallel network for in-the-wild facial expression recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2410–2414.

[11] Z. Xing, W. Tan, R. He, Y. Lin, and B. Yan, "Co-completion for occluded facial expression recognition," in *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, J. Magalhães, A. D. Bimbo, S. Satoh, N. Sebe, X. Alameda-Pineda, Q. Jin, V. Oria, and L. Toni, Eds. ACM, 2022, pp. 130–140.

[12] 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.

[13] 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.

[14] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Transactions on Affective Computing*, no. 01, pp. 1–1, 2021.

[15] Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 4, 2021, pp. 3510–3519.

[16] C. Wang, J. Xue, K. Lu, and Y. Yan, "Light attention embedding for facial expression recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 1834–1847, 2022.

[17] G. Yovel and B. Duchaine, "Specialized face perception mechanisms extract both part and spacing information: Evidence from developmental prosopagnosia," *Journal of Cognitive Neuroscience*, vol. 18, no. 4, pp. 580–593, 2006.

[18] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[19] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. a. Gelly, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[21] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[22] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6023–6032.

[23] Y. Wang, V. I. Morariu, and L. S. Davis, "Learning a discriminative filter bank within a cnn for fine-grained recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4148–4157.

[24] L. Zhang, S. Huang, W. Liu, and D. Tao, "Learning a mixture of granularity-specific experts for fine-grained categorization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8331–8340.

[25] S. Huang, X. Wang, and D. Tao, "Stochastic partial swap: Enhanced model generalization and interpretability for fine-grained recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 620–629.

[26] 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.

[27] L. A. Halliday, "Emotion detection: can perceivers identify an emotion from limited information?" Master's thesis, University of Canterbury Psychology, 2008.

[28] M. Nusseck, D. W. Cunningham, C. Wallraven, and H. H. Bülthoff, "The contribution of different facial regions to the recognition of conversational expressions," *Journal of vision*, vol. 8, no. 8, pp. 1–1, 2008.

[29] 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.

[30] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European conference on computer vision*. Springer, 2016, pp. 87–102.

[31] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 4690–4699.

[32] N. Ma, X. Zhang, H. Zheng, and J. Sun, "Shufflenet V2: practical guidelines for efficient CNN architecture design," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11218. Springer, 2018, pp. 122–138.

[33] 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.

[34] 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.

[35] Y. Zhang, C. Wang, and W. Deng, "Relative uncertainty learning for facial expression recognition," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 616–17 627, 2021.

[36] 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.

[37] D. Ruan, Y. Yan, S. Lai, Z. Chai, C. Shen, and H. Wang, "Feature decomposition and reconstruction learning for effective facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7660–7669.

[38] F. Xue, Q. Wang, and G. Guo, "Transfer: Learning relation-aware facial expression representations with transformers," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 3581–3590.

[39] D. Zeng, Z. Lin, X. Yan, Y. Liu, F. Wang, and B. Tang, "Face2exp: Combating data biases for facial expression recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 20 259–20 268.

[40] N. Le, K. Nguyen, Q. D. Tran, E. Tjiputra, B. Le, and A. Nguyen, "Uncertainty-aware label distribution learning for facial expression recognition," in *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*. IEEE, 2023, pp. 6077–6086.

[41] Y. Zhang, C. Wang, X. Ling, and W. Deng, "Learn from all: Erasing attention consistency for noisy label facial expression recognition," in *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVI*, ser. Lecture Notes in Computer Science, S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., vol. 13686. Springer, 2022, pp. 418–434.

[42] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[43] 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 *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

Yansong Huang obtained the Master's degree in the School of Computer Engineering and Science, Shanghai University, Shanghai, China. His research interests include computer vision and deep learning.

Junjie Peng obtained all of his degrees from bachelor to doctorate from Harbin Institute of Technology, China. He is a professor with the School of Computer Engineering and Science, Shanghai University, Shanghai, China. His research interests cover natural language processing, sentiment analysis, computer vision, intelligent technology and systems, data analysis and visualization etc. He is an author of more than 140 referred papers presented in international journals and conference proceedings, an owner of more than thirty patents. He is a senior member of CCF and a member of IEEE.

Wenqiang Zhang received the BS degree from Huazhong University of Science and Technology, China, in 1992, the MS degree from Shandong University, China, in 2000, the PhD degree from Shanghai Jiao Tong University, Shanghai, China, in 2004. He is a professor with the School of Computer Science, Fudan University, Shanghai, China. His current research interests include robotics, computer vision, machine intelligence, intelligent equipment, video/image analysis, etc. He has been engaged in the production, education, and research of robotics, artificial intelligence, medical informatization and intelligent equipment for a long time. He is an author of more than 100 papers and has applied for more than 50 patents (20 patents have been authorized). He is a senior member of CCF and a senior member of CSIG.

Tong Zhao obtained the Master's degree in the School of Computer Engineering and Science, Shanghai University, Shanghai, China. Her research interests include multimodal sentiment analysis and deep learning.

Gan Chen received the B.E degree in computer science from Nanchang Hangkong University, Jiangxi, China, in 2004, and the M.S degree in computer science from Jiangxi University of Science and Technology, Jiangxi, China, in 2010. She is currently pursuing the PH.D. degreee with the School of Computer Engineering and Science, Shanghai university. Her research interests include computer vision and deep learning.

Shuhua Tan is a deputy director of YTO Express Co., Ltd. and the national logistics engineering laboratory, a outstanding academic member of CCF YOCSEF, Shanghai, China. He joined Yuantong express in 2009 and is responsible for the planning, architecture design, and R & D management and science and technology projects of National Laboratory and innovation.

Fen Yi is an innovator and R & D expert of YTO Express Co., Ltd. and the national logistics engineering laboratory, Shanghai, China. She joined YTO Express in 2010 and is responsible for the innovation and application of express information technology.

PLACE PHOTO HERE

Lu Wang received the BS degree from Xi'an Jiaotong University, China, in 2002, the PhD degree from Tsinghua University, China, in 2008. He is an assistant professor with the School of Computer Engineering and Science, Shanghai University, Shanghai, China. His research interests include machine learning, pattern recognition, computer vision, etc.