CF-DAN: Facial-Expression Recognition Based on Cross-Fusion Dual-Attention Network

Fan Zhang
Shandong Technology and Business University
Shandong 264005, China
zhangfan@sdtbu.edu.cn

Gongguan Chen
Shandong Technology and Business University
Shandong 264005, China
2020410018@sdtbu.edu.cn

Hua Wang
Ludong University
Shandong 264025, China
hwa229@163.com

Caiming Zhang
Shangdong University
Shandong 250100, China
czhang@sdu.edu.cn

Abstract

In recent times, facial-expression recognition (FER) primarily focuses on images in the wild that include factors, such as face occlusion and image blur, rather than on laboratory images. The complex field environment has brought new challenges to FER. To address these challenges, this paper proposes a cross-fusion dual-attention network. The network comprises three parts—a cross-fusion grouped dual-attention mechanism to refine local features and obtain global information; a C^2 activation function construction method is proposed, and the new activation function is a piecewise cubic polynomial with three degrees of freedom, it not only requires less computation, but also has better flexibility and recognition ability, which can better solve the problems of slow running speed and neuron inactivation; and a closed-loop operation between the self-attention distillation process and residual connections to suppress redundant information and improve the generalization ability of the model. The recognition accuracy on the RAF-DB, FERPlus and AffectNet datasets were 92.78, 92.02, and 63.58%, respectively. Experimental results showed that this model could provide a more effective solution for FER tasks.

***Keywords:** Facial expression recognition, Cubic polynomial activation function, Dual-attention mechanism, Interactive learning, Self-attentional distillation.*

1. Introduction

The study of human emotional states is an interdisciplinary research field spanning psychology and computer science, and it is a fundamental undertaking in developing emotional intelligence. Facial expressions are the most natural and powerful external expression of a person's emotional state, such as their calmness, happiness, anger, sadness, fear, disgust, and surprise, and are key to human nonverbal communication. With the continuing technological developments, the research on facial-expression recognition (FER) has deepened, which has had a major impact on many facets of life, such as public security, lie detection, driving-fatigue detection, intelligent medical treatment[11], and security monitoring, amongst others[20, 38].

Traditional FER primarily used manual features and shallow-learning methods[44, 33], directional gradient[36], sparse representation[42], and non-negative matrix decomposition[1]. In recent years, with the wide application of deep learning in computer vision, convolutional neural networks (CNNs)[4, 24, 51, 29] have almost completely replaced traditional methods and achieved excellent results in image classification, attitude estimation[47], attribute learning, image segmentation, and other fields. Compared with traditional facial recognition methods, deep learning can extract deeper features and is more flexible in abstract representation of images. With the emergence of several better-performing networks—such as residual networks (ResNets) and recurrent neural networks (RNNs)—in various computer vision tasks, many

researchers have applied these networks to FER tasks, making excellent progress[13, 12, 50, 30].

In recent studies, the self-attention mechanism has been widely used in various computer vision tasks[26, 21, 41, 49]. It can imitate how people pay attention to the key positions within an image and extract key information from these key positions to complete various tasks. For example, the vision transformer (ViT)[10] model can be applied to image classification. However, in the FER field, networks that are similar to the ViT model cannot be directly used owing to the limited sample size. Aouayeb et al.[3] successfully transferred the ViT network into the expression recognition task by adding the squeeze and excitation (SE) block. The FER-VT[17] method sends feature maps of different scales into the same Transformer model to complete information fusion. Although the above methods expand the size of the operation window, they also introduce too many invalid connections, which reduces the model's ability to obtain global information. The dual attention Transformer network[31, 37, 9, 14] complements each other in feature refinement and global information acquisition. The parallel computing strategy used in this paper not only embodies the powerful feature extraction ability, but also avoids the massive information loss caused by serial computing. Moreover, we designed a cross-fusion feature extraction module based on the dual-attention mechanism to better complete the information interaction in different dimensions. To sum up, the contributions of this paper are as follows:

(1) We proposed a cross-fusion dual-attention Transformer based on spatial dimension and channel dimension. Local interaction in the spatial dimension completes feature refinement, and a global receptive field is provided in the channel dimension. The cross-fusion dual-attention Transformer realizes the mutual complementation of the feature information of different dimensions, thereby improving the accuracy of the respective features.

(2) We designed a construction method of activation functions to solve the problems in commonly used activation functions, such as neuron deactivation, the large computational overhead brought about by the power function, and the existence of non-differentiable points. At the same time, a new C^2 continuous activation function is constructed for the interactive learning mechanism in this paper, which improves the ability of the interactive learning mechanism to integrate different features.

(3) To reduce the high computational cost caused by the self-attention mechanism, we proposed a grouped mechanism and a self-attention distillation to act on the self-attention mechanism. This process divided attention into different groups, and used the self-attention distillation in each group to reduce the spatial dimension of K and V, so as to reduce the computational cost. Using self-attention distillation not only improves the ability of the self-attention mechanism, but also significantly reduces the computational cost (33%).

2. Related work

FER in the real-world environment: To address the challenges of FER in the real-world environment, researchers have made great efforts to improve recognition accuracy under complex background and figure occlusion conditions. Bourel et al.[6] extracted the spatial degree features of key parts of the face for classification. PCA-based methods[45, 18], realized the projection of test images to training images, using similarity to realize expression recognition. Hammal et al.[15] used image multi-level segmentation and sparse decomposition to solve the facial-expression occlusion problem. Multiple face detectors, such as the MTCNN[48] and Dlib[2] models, have been used for facial detection in real-world scenes. These detection algorithms can be used to preprocess the expression images, after which the detected face regions are sent to various models for expression recognition. Most researchers have devised occlusion-aware-based methods to remove the distractions brought about by complex backgrounds. Moreover, to solve the problems created by real-world scenarios, more and more multiview- and multiscale-based research have been conducted. Happy[16] and Majumder[28] proposed that facial expression changes were mainly reflected in key parts, such as the eyes and mouth. Subsequently, researchers increasingly focused on how to use key parts to extract key information to increase FER accuracy.

Visual Transformers: Several recent studies have shown that Transformers have enormous potential in computer vision applications. Several pioneering studies such as those on the iGPT[7] and ViT methods applied the self-attention mechanism directly to image pixel or patch sequences. Inspired by this, convolutional visual transformers (CVTs)[27] were the first to apply the Transformer model to FER tasks. The CVT uses the local binary patterns (LBP) algorithm to send the facial expression images in two different states into the ResNet network to obtain smaller feature images, with the Transformer model being used to complete FER. The mask vision transformer (MViT)[22] generates a mask based on the Transformer model for filtering complex backgrounds and the occlusion of facial images.

3. Methodology

To better solve the FER task in real-world environments, a more concise and effective dual-attention mechanism is proposed . Considering the large size of the feature map will seriously increase the computation of the self-attention mechanism, the network first uses the ResNet model for high-dimensional mapping of facial expression images, thus acquiring high-dimensional feature maps of a smaller scale.

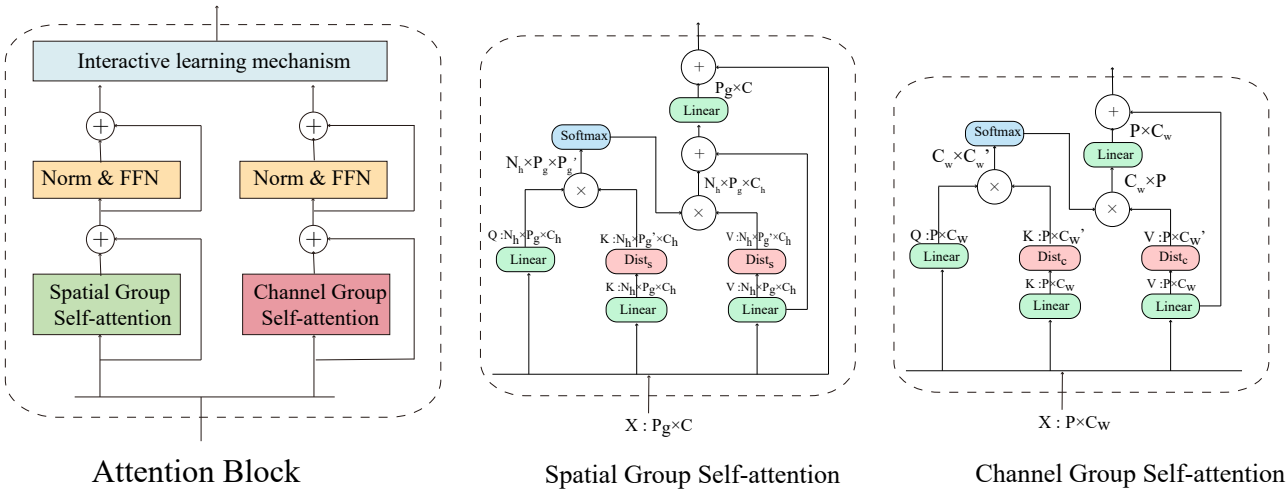


Figure 1. Dual-attention mechanism.

The overall flow of the network is depicted in Figure 1 and Figure 2. The left side of Figure 1 depicts the overall structure of the dual attention mechanism. Self-attention is used to realize facial expression recognition, which solves the problem of small receptive field of convolutional neural network. The introduction of self-attention mechanism with different dimensions makes up for the deficiency of traditional single dimension in global interaction. The parallel algorithm directly applies double attention to the original input, overcomes the defect that self-attention in the channel dimension only enhances the result of spatial dimension self-attention, and makes the model more sensitive to the surrounding environment. Compared with the traditional self-attention mechanism, parallel dual attention can be better applied to facial expression recognition in real environment. The right side of Figure 1 depicts the flow of self-attention in different dimensions. Input Q, K and V are obtained by linear variation and self-attention distillation, and then self-attention is calculated. In the face of the problem that expression data is difficult to collect, the grouping method and the addition of self attention distillation avoid using the GAN network to generate data, greatly reduce the calculation cost of the model, and improve the problem that the traditional self attention mechanism can not work well on small data sets.

Figure 2 is the overall structure of the cross-fusion dual attention network in this paper. The main part of the model is composed of two layers of dual attention and an interactive learning mechanism. The information sharing is accomplished by exchanging K between the two layers of dual attention, breaking the information occlusion caused by parallel operations. In order to improve the feature fusion ability of the interactive learning mechanism for dual attention, we design a C^2 continuous activation function $PCP(x)$ composed of three-segment cubic polynomials, which has better feature fusion ability. Compared with other commonly used activation functions in the ablation experiment, it is proved that $PCP(x)$ can achieve higher recognition accuracy in the expression recognition task.

3.1. Dual-attention mechanism(DAM)

With the widespread application of sparse attention mechanisms, most researchers have chosen to reconstruct the input image into the form of $R^{P\times C}$, by dividing the spatial dimension into patches and adding positional coding, where P denotes the number of patches and C denotes the number of channels. However, the unique two-dimensional information of images can be destroyed, so the refinement of local features becomes an important challenge. To address these problems, this paper divides the reconstructed images into groups, conducting a separate self-attention mechanism in each group. Consider that the number of channels is not corrupted during refactoring, and that each channel is an abstract representation of the global information. In this study, the self-attention mechanism of the channel dimension and the self-attention mechanism of the space dimension are operated in parallel to form a dual-attention mechanism. Similarly, group-based learning is added to the channel dimension. The dual-attention mechanisms complement each other in the acquisition of local features and global information, exhibiting strong FER abilities.

Specifically, in the spatial dimension, it can be assumed that the reconstructed image is divided into N_g groups, each group containing P_g patches. The operational process of the overall self-attention mechanism in the spatial dimension

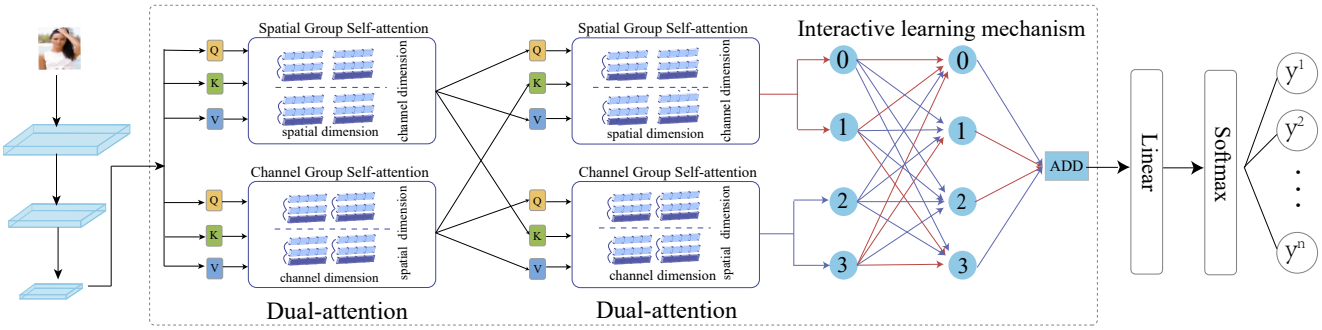


Figure 2. Overall structure of model.

can be expressed as follows:

$$A_s(Q,K,V) = Concat(A_1^*(Q_1,K_1,V_1),...,A_{N_g}^*(Q_{N_g},K_{N_g},V_{N_g})) \quad (1)$$

where $Q,K,V \in \mathbb{R}^{P\times C}$ can be obtained by a linear transformation of the input. $A_1^*(Q_1,K_1,V_1),...,A_{N_g}^*(Q_{N_g},K_{N_g},V_{N_g})$ represents the result of self-attention, $Q_1,K_1,V_1...Q_{N_g},K_{N_g},V_{N_g} \in \mathbb{R}^{P_g\times C}$.

In the channel dimension, it can be assumed that the reconstructed image is divided into N_w groups, each group containing C_w channels, so that $C = C_w \times N_w$. The operational process of the overall self-attention mechanism in the channel dimension can be expressed as follows:

$$A_c(Q,K,V) = Concat(A_1^*(Q_1,K_1,V_1),...,A_{N_w}^*(Q_{N_g},K_{N_g},V_{N_g})) \quad (2)$$

where $Q,K,V \in \mathbb{R}^{P\times C}$ can be obtained by a linear transformation of the input. $A_1^*(Q_1,K_1,V_1)$, ..., $A_{N_w}^*(Q_{N_g},K_{N_g},V_{N_g})$ represents the result of self-attention, $Q_1,K_1,V_1...Q_{N_g},K_{N_g},V_{N_g} \in \mathbb{R}^{P\times C_w}$.

Equations (1)-(2) illustrate that all spatial locations are taken into account when calculating channel-dimensional self-attention, which enables it to have the ability to interact globally. In subsequent ablation studies, we also confirmed that channel-dimensional self-attention pays more attention to the face as a whole and the connection of facial expressions to the surrounding environment. The spatial dimension of self-attention is limited to different spatial locations to complete local interactions, which makes it more sensitive to key locations such as eyes and mouth. The two cooperate with each other to enhance the perception of facial expressions. At the same time, in order to realize the timely sharing of information when the two dimensions are processed in parallel, we use the crossover method to complete the information transfer between different dimensions. We will go into details in the next subsection.

3.2. Cross-fusion attention mechanism

Expression images from real-world environments can be complex and diverse. People express the same emotion very differently. Moreover, there are many similarities in the expression of different emotions. Consequently, the simple application of neural networks to facial expression images cannot accurately distinguish minute nuances, resulting in low recognition rates. In the dual self-attention mechanism, self-attention of the spatial dimension pays more attention to the key regions related to facial expression. Contrastingly, self-attention of the channel dimension pays more attention to global information. Both types of information are important for FER. To make better use of information from different dimensions, we designed the cross-fusion dual self-attention model as shown in Figure 2. In the previous research of cross fusion, additional data support is often needed to complete the cross between features generated by different data. In this paper, we adopt a more concise implementation method, which directly adds cross-fusion in the operation process of two different dimensions of self-attention mechanism. Experiments show that our design is effective.

Specifically, the output $A_s^l \in \mathbb{R}^{P\times C}$ of the upper layer of the spatial dimension is mapped to two image matrixes, Q_s^l and V_s^l, through two linear transformations, and the output $A_c^l \in \mathbb{R}^{P\times C}$ of the upper layer of the channel dimension is mapped to the image matrix K_c^l through a linear transformation, before sending Q_s^l, K_c^l and V_s^l into the spatial dimension of the next layer for self-attention. The operation is similar in the channel dimension, the attention result of which can be expressed as follows:

$$\begin{aligned} A_s^{l+1} &= A_s(Q_s^l,K_c^l,V_s^l) \\ A_c^{l+1} &= A_c(Q_c^l,K_s^l,V_c^l) \end{aligned} \quad (3)$$

3.3. Self-attentional distillation

The self-attention mechanism expands the attention window to the whole image, greatly increases the compu-

tational overhead, and produces severe smearing phenomenon, causing many redundant combinations in K and V. Excessive useless information introduces serious interference to the feature extraction process, resulting in declining model performance. This paper proposes that the self-attention distillation mechanism acts on keys and values, achieving a reduction of scale in the spatial and channel dimensions. Dominant features can be extracted using this operation to form a feature map with a focused advantage in subsequent self-attention, suppressing interference from redundant information and reducing noise generation. Moreover, to reduce the loss of middle- and high-frequency information caused by the self-attention distillation mechanism, a residual connection can be used to fuse the original V with the results of the self-attention process. The interaction of self-attention distillation and residual connection constructs an independent closed-loop operation, which effectively reconstructs the lost information.

Specifically, in terms of the spatial dimension, two convolution layers with a kernel size of three are constructed. The first layer completes the mapping of the channel number from a high dimension to a low dimension, which is a dynamic process to realize the extraction of dominant features. The second layer completes the mapping of the channel number from a low dimension to a high dimension, thus maintaining the same dimension as the original input. Finally, max pooling is used to reduce the number of patches in the keys and values. The overall calculation process can be expressed as follows:

$$\begin{aligned} Dist_s(K) &= MaxPool(Conv_{3+}(Conv_{3-}(K))) \\ Dist_s(V) &= MaxPool(Conv_{3+}(Conv_{3-}(V))) \end{aligned} \tag{4}$$

where $Conv_{3-}$ denotes the convolution operation with a convolution kernel size of three (used to reduce the number of channels), $Conv_{3+}$ denotes the convolution operation with a convolution kernel size of three (used to increase the number of channels), and $MaxPool$ denotes the max pooling.

In terms of the channel dimension, the convolution operation with a convolution kernel size of one is used to reduce the number of channels in a single group, after which the convolution operation with a convolution kernel size of three is used to enhance the connection between channels to learn more reliable high-quality features. To guide the model to focus on the acquisition of global information, max pooling is abandoned in the channel dimension so that the number of patches in each group remains unchanged. The overall calculation process can be expressed as follows:

$$\begin{aligned} Dist_c(K) &= Conv_3(Conv_1(K)) \\ Dist_c(V) &= Conv_3(Conv_1(V)) \end{aligned} \tag{5}$$

where $Conv_1$ and $Conv_3$ denote convolution operations with convolution kernel sizes of 1 and 3, respectively.

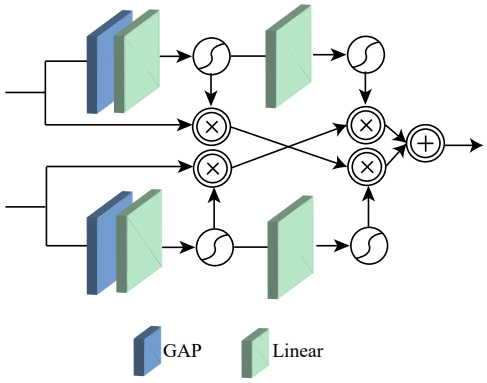


Figure 3. Interactive learning mechanism.

3.4. Interactive learning mechanism(ILM)

Because activation functions, such as Sigmoid and Tanh, have exponentiation operations, the calculation speed slows, and the ReLu activation function has non-differentiable points and neuron-death problems. To solve these problems, We designed a C^2 continuous activation function composed of three-segment cubic polynomial curves, called piecewise cubic polynomial function, or $PCP(x)$ for short, so as to increase the ability of the interactive learning mechanism to feature fusion. The construction method of $PCP(x)$ is as follows.

First, the interval [-1,1] is divided into three intervals by points $x_1 = -1$, $x_2 = 0$, $x_3 = x_L$ and $x_4 = 1$, and the corresponding function values on the four points x_1, x_2, x_3, x_4 are P_1, $P_2 = 0$, P_3, $P_4 = 1$ respectively , the first derivative at point x_2 is defined by:

$$\frac{dP_2}{dx} = s_L \frac{P_2 - P_1}{x_2 - x_1} \tag{6}$$

where x_L is a pending parameter.

The first derivatives at points x_1 and x_4 are both 0, $\frac{dP_1}{dx} = \frac{dP_4}{dx} = 0$. In this way, we can construct the Hermite cubic interpolation function in the interval $[x_i, x_{i+1}], i = 1, 2, 3$:

$$P_i(x) = F_0(x)F_i + G_0(x)\frac{dP_i}{dx} + F_1(x)F_{i+1} + G_1(x)\frac{dP_{i+1}}{dx} \tag{7}$$

where

$$\begin{aligned} F_0(x) &= (x_{i+1} - x)^2 \left(2\left(x - x_i\right) + h\right)/h^3 \\ F_1(x) &= (x - x_i)^2(2(x_{i+1} - x) + h)/h^3 \\ G_0(x) &= (x_{i+1} - x)^2(x - x_i)/h^2 \\ G_1(x) &= -(x - x_i)^2(x_{i+1} - x)/h^2 \\ h &= x_{i+1} - x_i \end{aligned} \tag{8}$$

There are three undetermined parameters x_L, P_1 and s_L in equation (8). When the values of parameters x_L, P_1 and

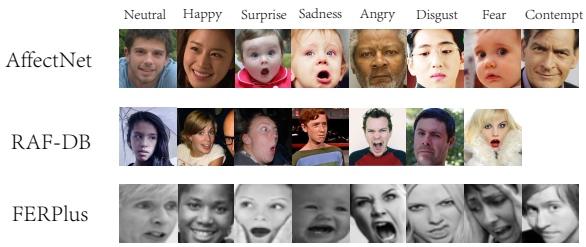


Figure 4. Partial samples from three data sets.

s_L are given, P_3 and $\frac{dP_3}{dx}(x_3)$ in equation 8 can be calculated by the following method. To ensure that $PCP(x)$ is C^2 continuous, the second derivative at point x_2 is defined by $\frac{d^2P_1}{dx^2}(x_2)$, such that P_3 and P_3 and $\frac{dP_3}{dx}(x_3)$ can be calculated by the following system of equations:

$$\begin{aligned} \frac{d^2P_2}{dx^2}(x_2) &= \frac{d^2P_1}{dx^2}(x_2) \\ \frac{d^2P_2}{dx^2}(x_3) &= \frac{d^2P_3}{dx^2}(x_3) \end{aligned} \tag{9}$$

From the above construction process, $PCP(x)$ is C^2 continuous. The definition of $PCP(x)$ is as follows:

$$PCP(x) = \begin{cases} P_1 & \text{if } x < -1 \\ P_1(x) & \text{if } -1 \leq x < 0 \\ P_2(x) & \text{if } 0 \leq x < x_L \\ P_3(x) & \text{if } x_L \leq x < 1 \\ 1 & \text{if } x > 1 \end{cases} \tag{10}$$

Properly adjusting the values of the three parameters x_L, P_1 and s_L in $PCP(x)$ can improve the accuracy of the algorithm, thereby enhancing the ability of the interactive learning mechanism to feature fusion.

Note: Since $P_i(x)(8)$ is applied repeatedly in the model, it takes too much computational cost to directly calculate $P_i(x)$ using Equation 8. For this reason, we write $P_i(x)(12)$ in the following form:

$$P_i(x) = ax^3 + bx^2 + cx + d \tag{11}$$

where a, b, c and d are the coefficients obtained by simplifying Equation 8, respectively.

Calculate a, b, c and d in $P_i(x)(8)$ before each iteration, so that $P_i(x)(8)$ can be calculated using the following formula:

$$P_i(x) = ((ax + b)x + c)x + d \tag{12}$$

In this way, only 3 multiplications and 3 additions are required to calculate $P_i(x)$, which significantly reduces the amount of computation. The calculation by Equations 8 and 9 requires 4 divisions, 17 multiplications, 4 subtractions and 3 additions.

On the basis of $PCP(x)$, this paper proposes an interactive learning mechanism, as shown in Figure 3. The global average pooling of individual feature vectors in the channel dimension is first used to integrate the global spatial information, after which the feature vector is mapped through a feedforward neural network. We refer to the mapping results as implicit dynamic vectors. Finally, the eigenvector and implicit dynamic vector are cross-multiplied to complete interactive learning. The global average pooling uses the mean value of the feature map to forcibly demarcate its importance, directly giving each channel its actual meaning, and then we feed the result of the global average pooling to the input in the form of weights to enhance the input features. In the interaction process, since the activation function can compress the negative area to a smaller negative interval, the interactive information can be identified, and the non-interactive information can be suppressed to an inactive state, so that it can better integrate with other dimensions of information.

Specifically, the generation process of implicit dynamic vectors and the overall operational process of the interactive learning mechanism can be expressed as follows:

$$CA(X) = PCP(Linear(GAP(X))) \tag{13}$$

$$CAM(X) = CA(X) \times X \tag{14}$$

$$\begin{aligned} &INTA(X_A, X_B) = \\ &\quad PCP(Linear(CA(X_A))) \times CBM(X_B) \end{aligned} \tag{15}$$

$$\begin{aligned} &ILM(A_s(Q,K,V), A_c(Q,K,V)) = \\ &\quad INTA(A_s(Q,K,V), A_c(Q,K,V)) + \\ &\quad INTA(A_c(Q,K,V), A_s(Q,K,V)) \end{aligned} \tag{16}$$

where PCP denotes the activation function, $Linear$ denotes the full connection layer, and GAP denotes the global average pooling.

4. Experimental Results

This section includes the dataset, experimental environment, and experiment implementation details. Some samples are shown in Figure 4. The effectiveness of the proposed model is compared with commonly used methods of recent years. Subsequently, the improvement of each part of the model is investigated using ablation experiments.

4.1. Datasets

We evaluated our method on three commonly used facial expression datasets: the AffectNet, RAF-DB, and FERPlus datasets, which were collected in a real-world environment, subject to different degrees of light and occlusion.

Methods	Year	RAF-DB	AffectNet	FERPlus
SCN[39]	CVPR 2020	87.03	60.23	89.39
RAN[40]	TIP 2020	86.90	-	89.16
EfficientFace[52]	AAAI 2021	88.36	59.89	-
DMUE[34]	CVPR 2021	89.42	-	-
FDRL[32]	CVPR 2021	89.47	-	-
DAN[43]	arXiv 2021	89.70	62.09	-
ARM[35]	arXiv 2021	90.42	61.33	-
CSGResNet[19]	ICASSP 2022	88.59	61.03	88.94
AMP-Net[25]	TCSVT 2022	89.19	61.32	89.37
POSTER[53]	arXiv 2022	92.05	63.34	91.62
Ours	-	92.78	63.58	92.02

Table 1. Comparison on RAF-DB, AffectNet, and FERPlus datasets

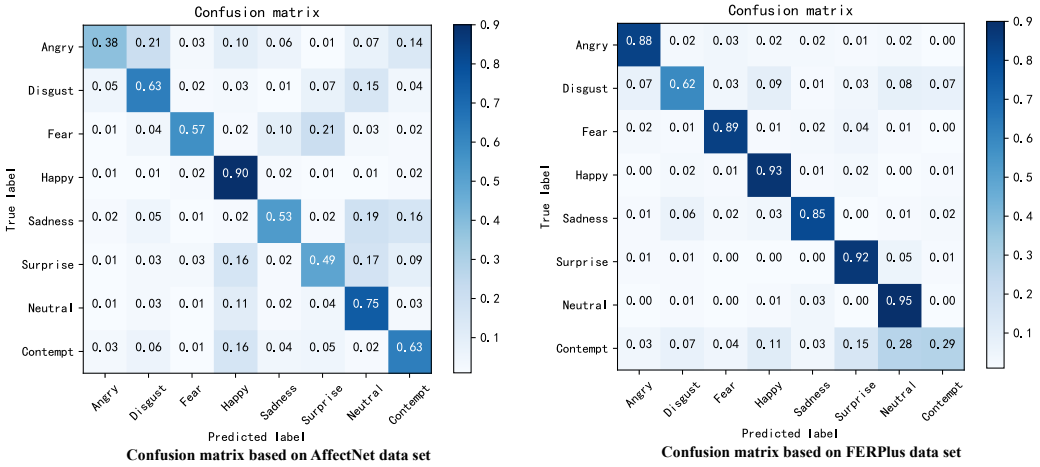


Figure 5. Confusion matrix based on AffectNet and FERPlus.

AffectNet[8][1]: AffectNet is a large outdoor facial expression dataset comprising over a million facial images from the Internet. The dataset contains eight categories. It is important to note that AffectNet's training and test sets are extremely unbalanced.

RAF-DB[23][2]: RAF-DB is a dataset of facial expressions in real life scenes. The dataset comprises seven categories. The training set contains 12,271 samples, and the test set contains 3,068 samples.

FERPlus[5][3]: FERPlus relabeled mislabeled images and removed non-face images from the original FER2013 dataset. Each image in FERPlus had multiple annotators participating, providing better tag quality than the original FER2013 dataset. Like AffectNet, the dataset has a total of eight categories.

[1]http://mohammadmahoor.com/affectnet/
[2]http://www.whdeng.cn/raf/model1.html
[3]https://www.worldlink.com.cn/osdir/ferplus.html

4.2. Experimental Details

The model was implemented using Python 3.7 and Pytorch 1.7.1. For all training cases, face images were detected using the MTCNN network. During the experiment, all images were further sized to 224 ×224 pixels. The model was trained on a single NVIDIA GTX 2080 GPU graphics card. During the training, the batch size was set to 32, and the AdamW optimizer with a momentum of 0.9 and weight attenuation of 1e-4 was used to optimize the model. During the training process, the model only used the cross-entropy loss function, giving it good generalization ability.

4.3. Results and Analysis

Here, we compare the proposed model with the most advanced methods used in recent years on the AffectNet, RAF-DB, and FERPlus datasets, to prove the superiority of our method in FER tasks. The comparative results are shown in Table 1.

On the RAF-DB dataset, compared with the Efficient-

Methods	Year	Params	FLOPs	Acc(RAF-DB)	Acc(FERPlus)
CVT	arXiv 2021	80.1M	-	87.61	88.81
DMUE	CVPR 2021	78.4M	13.4G	89.42	-
TransFER	ICCV 2021	65.2M	15.3G	90.91	90.83
POSTER	arXiv 2022	71.8M	15.7G	92.05	91.62
DAN	arXiv 2021	28.3M	2.6G	89.70	-
Ours	-	16.4M	2.0G	92.78	92.02

Table 2. Comparison on Parameters and FLOPs

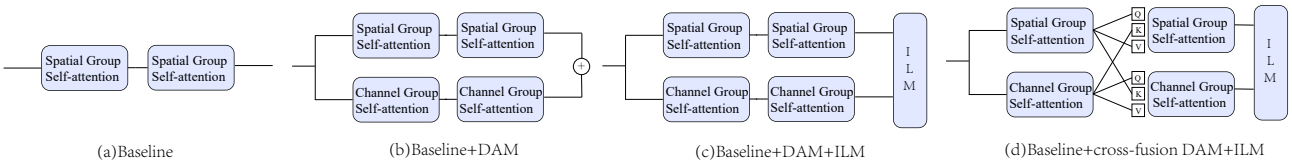


Figure 6. Ablation studies of dual-attention and interactive learning mechanism.

Face, SCN, and other CNN-based methods, the model proposed in this paper exhibits an improvement of approximately 4%. Compared with the DAN, AMP-Net, and other attention-based methods, the improvement is approximately 2%. Experiments demonstrate that our proposed model has more advanced recognition capabilities on the RAF-DB dataset and provides a more effective solution for FER tasks. For the AffectNet dataset, based on the data shown in the table, the identification accuracy of our proposed method is 63.58%. Compared with the SCN, EfficientFace, and RAN methods, the improvement rate is approximately 4%, and compared with the DMUE, FDRL, DAN, ARM, CSGResNet, and AMP-Net methods, the improvement rate is approximately 2%. For the FERPlus dataset, the improvement rate is over 1% compared to the other methods.

To explore the performance of this model more accurately under different facial expressions, we examined the confusion matrices in the two datasets, as shown in Figure 5. The confusion matrices describe in detail the recognition accuracy of each expression and the proportion misclassified as other expressions, where the diagonal item represents the recognition accuracy of each expression. As is evident from the data, the Happy expression is the easiest to recognize among the eight expressions owing to its large display range. The recognition rate of Happy in the AffectNet dataset is considerably higher than that of the other expressions. Apart from Happy faces, the difference in success rate was small, primarily because the images in the AffectNet dataset came from the Internet and contained many error samples. In the FERPlus dataset, the Happy recognition rate is slightly lower than that of Neutral because Neutral expressions have the largest sample size. Additionally, the Disgust and Contempt samples in the FERPlus dataset number the least—just one tenth of the number of samples of other expressions—and these expressions have similar appearance characteristics. Consequently, the recognition accuracy of Disgust and Contempt is substantially lower than that of the other expressions.

In the recent studies on FER, the Transformer has received increasing attention. However, the enormous number of parameters remains a major limitation in using the Transformer. Moreover, model parameters (Params) and floating-point operations (FLOPs) are also two key features to be considered for fair comparisons. One of the starting points of this study was to reduce the overall number of model parameters. To solve this problem, we proposed a grouped self-attention mechanism and a self-attention distillation mechanism. Table 2 shows a comparison of the parameters of the proposed method and other methods, including the CVT, DMUE, TransFER[46], POSTER, and DAN models. It is evident that the number of parameters in the proposed method is just a quarter of those for the other methods while maintaining the highest FER accuracy.

4.4. Ablation Study

Performance of dual-attention and interactive-learning mechanism: To verify that the dual-attention approach has considerable advantages, we first used a single self-attention mechanism as a baseline. The baseline was then compared with the dual-attention mechanism. We then compared the effects of interactive learning and cross-fusion dual attention on FER. The differences are shown in Figure 6(a)–6(d). The experimental results are shown in Table 3. It is evident that more advanced classification accuracy is achieved in FER tasks after the addition of channel dimension self-attention and interactive learning

Methods	AffectNet	FERPlus	RAF-DB
Baseline	59.73%	89.69%	90.37%
Baseline+DAM	60.44%	90.92%	90.88%
Baseline+DAM+ILM	63.12%	91.55%	91.63%
Baseline+cross-fusion DAM+ILM	63.58%	92.02%	92.78%

Table 3. Performance of dual attention and interactive learning mechanism

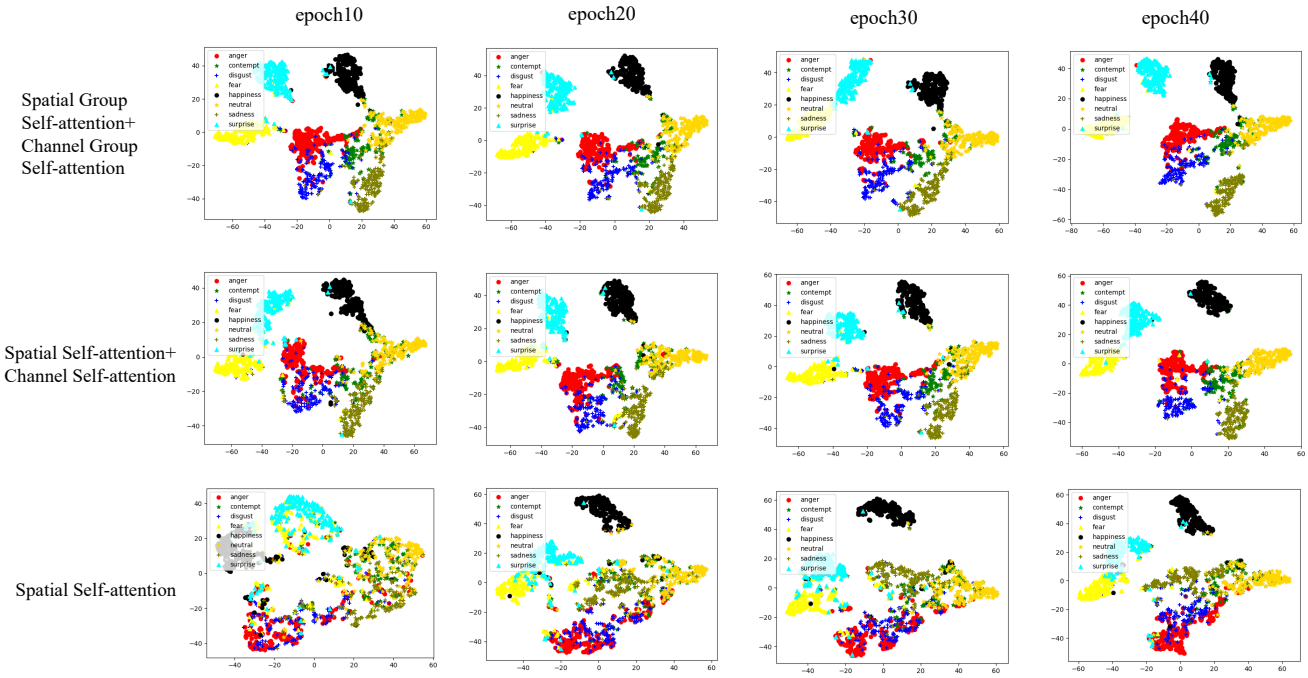


Figure 7. Two-dimensional feature map of some samples.

mechanism in parallel processing, proving the introduction of dual-attention mechanism to be feasible and effective. It can be seen from the third and fourth rows in Table 3 that adding interactive learning mechanism to FERPlus and RAF-DB datasets can bring about 1% improvement, and can bring about 3% improvement on AffectNet datasets. In order to better fuse the features from two different sources, we first consider the impact of one feature on the other, realize the mutual mapping of the two features, and complete the feature fusion on this basis, instead of simply adding or splicing. Experiments have proved that our idea is feasible. In addition, we also designed a new activation function construction method to make the interactive learning mechanism play a better role. In the later comparison of various activation functions, we can also prove the effectiveness of our proposed activation function.

To observe the effect of the dual-attention mechanism on feature extraction and the effect of grouping on FER, we used the t-SNE algorithm to visualize the expression features of part of the test set samples in two-dimensional space. The results are shown in Figure 7. It is evident that, because of the lack of global information, the recognition ability of the single-dimensional attention mechanism is considerably lower than that of the dual-attention mechanism in the initial training. The advantages of local key features emerge in the late training period. For the four expressions of Happy, Neutral, Surprise, and Fear, because of the large number of training samples and obvious facial features, they can be well recognized in the three conditions, which is consistent with the information obtained in the confusion matrix. For expressions with minor differences between them—such as Disgust and Contempt—the self-attention mechanism of the spatial dimension finds it difficult to separate them owing to the small number of training samples and the difficulty in distinguishing them. After the addition of the self-attention mechanism of the channel di-

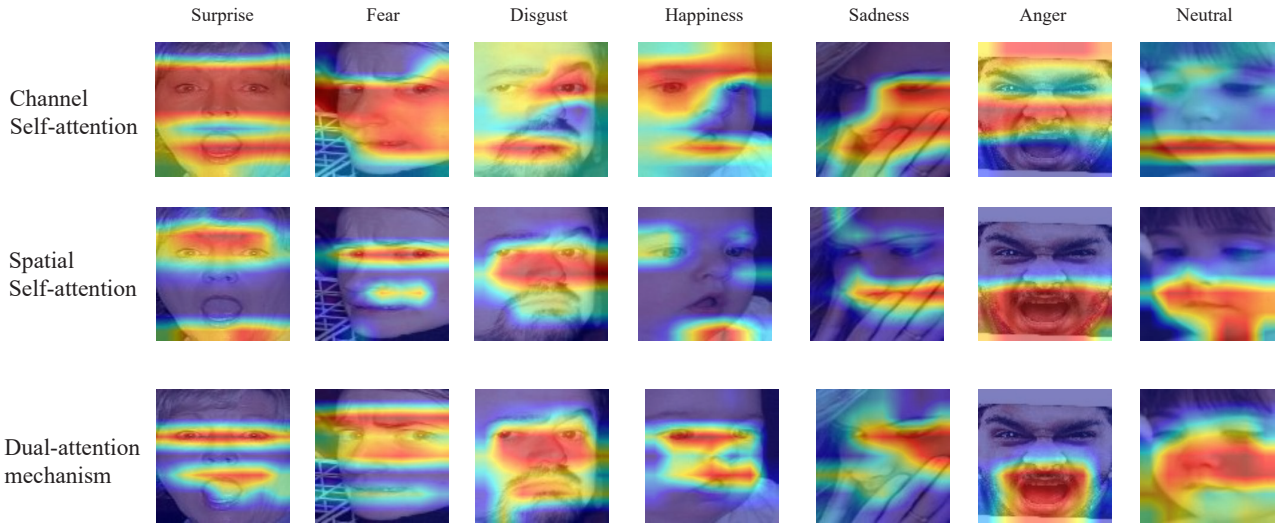
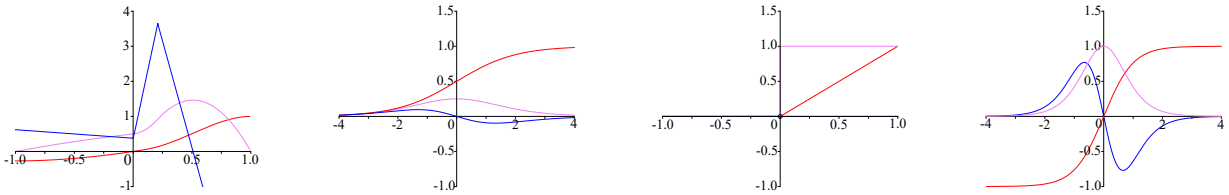


Figure 8. Attention visualization of the attention mechanism



(a) PCP (b) Sigmoid (c) ReLu (d) Tanh

Figure 9. Various activation functions and their derivatives. The red curve represents the activation function, the violet curve represents the first derivative, and the blue curve represents the second derivative

mension, there is still some overlap of facial features with small sample sizes, but it is substantially better than the single dimension self-attention mechanism. Moreover, it is evident that grouping in the dual-attention mechanism can effectively reduce the redundant connections caused by global attention and enhance the FER ability of the whole network. As is evident, the grouped dual-attention mechanism classifies Sadness, Contempt, and Disgust more accurately, with Sadness being the most obvious.

To study the effect of the fusion mechanism on the attention regions of facial expressions, we drew attention maps of the attention mechanism in the channel and space dimensions. As is evident from Figure 8, in the channel dimension the attention is focused on the overall facial expression and the interaction with the surrounding environment, while in the spatial dimension the attention is focused on key parts such as the eyes and mouth. The effect after fusion is the collection of key information from the two different dimensions.

Analysis of the validity of the activation function: As is evident from Figure 9, the Sigmoid function scales the value between 0–1, and the gradients in this interval are all less than 0.25, there being a potential hidden danger of gradient disappearance. The activation function proposed in this paper scales all values between a small negative number and 1, and the gradient in this interval is between 0–1.5. To solve the problem of neuron deactivation and discontinuous derivative in the negative part of the ReLu activation function, we mapped the negative region to a smaller negative interval. It is evident from Figure 9 that the activation function used in this paper is at all points continuously differentiable. Moreover, the proposed activation function does not involve exponentiation, improving the model's operational speed compared with that of the Sigmoid and Tanh activation functions. In terms of recognition accuracy, we compared various activation functions on the RAF-DB dataset, the experimental results of which are shown in Table 4. As is evident, the proposed activation function performs better in FER tasks.

Performance of self-attentional distillation: To verify

PCP	Sigmoid	ReLu	Tanh	ILM	ACC
✓					92.54
✓				✓	92.78
	✓				91.42
	✓			✓	91.78
		✓			91.98
		✓		✓	92.26
			✓		91.75
			✓	✓	91.97

Table 4. Performance of dual attention and interactive learning mechanism

Methods	AffectNet	FERPlus	Params	FLOPs
without distillation	63.34%	91.87%	4.2M	0.21G
with distillation	63.58%	92.02%	4.22M	0.14G

Table 5. Performance of self-attentional distillation

Figure 10. Parameter sensitivity analysis of activation functions

the effectiveness of self-attention distillation, the complete model was used as the benchmark, and then compared with the model discarded by self-attention distillation on the AffectNet and FERPlus datasets. The experimental results are shown in Table 5. It is evident that, after self-attention distillation is discarded, the operation window expands to the whole image, the serious trailing phenomenon leading to an increase of redundant information pairs, the introduction of excessive noise affecting the generalization of the whole model. When self-attention distillation and residual connection form a complete closed loop, information communication across windows is easier, while reducing the loss of middle- and high-frequency information. It can be concluded that self-attention distillation is worth introducing into the model. Moreover, it reduces the computational cost of the model.

4.5. Parameter sensitivity analysis

We performed a sensitivity analysis on variable parameters on the activation functions on the RAF-DB dataset. During the experiment, s_L was set to 1.75, 1.85, and 1.95 in three cases, x_L was set to 0.19, 0.2, and 0.21, and P_1 was set to -0.23, -0.24, -0.25, -0.26, and -0.27 in five cases. The experimental results are shown in Figure 10. As is evident from Figure 10, when s_L=1.75, the model performs poorly at x_L=0.19, and the performance is average in the other cases. At s_L=1.85, the model performs poorly at x_L=0.21, and when s_L=1.95 the model is relatively stable. When s_L=1.85, x_L=0.19, P_1=-0.24, the model achieves its highest accuracy.

5. Conclusion

In this paper, a cross-fusion dual attention network based on spatial dimension and channel dimension is proposed. The local interaction of the spatial dimension completes the feature refinement, and the channel dimension provides the global receptive field. The two kinds of self-attention features can be complemented by cross-fusion attention, so that the extracted features contain more effective information. The shape of the activation function has adaptive ad-

justment, which can increase the ability of feature extraction and fusion. It can also be applied to other learning frameworks to improve its accuracy and reduce the computational time. The method of constructing activation function proposed in this paper can also be extended to constructing polynomial activation function of $C^n(n > 2)$ according to the application, so as to increase the ability of reverse transfer of activation function and improve the generalization ability. Since the attention mechanism often require high computational cost, this paper puts forward the grouping mechanism and self-attention distillation act together on the self-attention mechanism. By dividing the attention into different groups, self-attention distillation is used in each group to reduce the spatial dimension of K and V, which improves the ability of the self-attention mechanism and reduces the computational cost.

Self-attention mechanism is one of the key technologies in facial expression recognition. In the following research, we will continue to explore how to construct more effective self-attention mechanism according to different data features in facial expression recognition task. At the same time, we will continue to study the relationship between self-attention mechanism and activation function. For specific data, we will try to construct more effective activation function adaptively, so that it has strong backpropagation ability, high continuity and low computational cost, so as to obtain stronger feature extraction ability with small computational cost.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 62272281 and 62007017, and the Youth Innovation Technology Project of Higher School in Shandong Province under Grant No. 2019KJN042.

References

[1] H. B. Ali, D. M. Powers, X. Jia, and Y. Zhang. Extended non-negative matrix factorization for face and facial expression recognition. *International Journal of Machine Learning and Computing*, 5(2):142, 2015. 1

[2] B. Amos, B. Ludwiczuk, M. Satyanarayanan, et al. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 6(2):20, 2016. 2

[3] M. Aouayeb, W. Hamidouche, C. Soladie, K. Kpalma, and R. Seguier. Learning vision transformer with squeeze and excitation for facial expression recognition. *arXiv preprint arXiv:2107.03107*, 2021. 2

[4] W. J. Baddar, S. Lee, and Y. M. Ro. On-the-fly facial expression prediction using lstm encoded appearance-suppressed dynamics. *IEEE Transactions on Affective Computing*, 2019. 1

[5] 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*, pages 279–283, 2016. 7

[6] F. Bourel, C. C. Chibelushi, and A. A. Low. Recognition of facial expressions in the presence of occlusion. In *BMVC*, pages 1–10. Citeseer, 2001. 2

[7] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 2

[8] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 2106–2112. IEEE, 2011. 7

[9] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang, and L. Yuan. Davit: Dual attention vision transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 74–92. Springer, 2022. 2

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[11] J. Edwards, H. J. Jackson, and P. E. Pattison. Emotion recognition via facial expression and affective prosody in schizophrenia. *Clinical psychology review*, 22(6):789–832, 2002. 1

[12] Y. Fan, X. Lu, D. Li, and Y. Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 445–450, 2016. 2

[13] M. Feffer, O. O. Rudovic, and R. W. Picard. A mixture of personalized experts for human affect estimation. In *International conference on machine learning and data mining in pattern recognition*, pages 316–330, 2018. 2

[14] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019. 2

[15] Z. Hammal, M. Arguin, and F. Gosselin. Comparing a novel model based on the transferable belief model with humans during the recognition of partially occluded facial expressions. *Journal of vision*, 9(2):1–19, 2009. 2

[16] S. Happy and A. Routray. Automatic facial expression recognition using features of salient facial patches. *IEEE transactions on Affective Computing*, 6(1):1–12, 2014. 2

[17] Q. Huang, C. Huang, X. Wang, and F. Jiang. Facial expression recognition with grid-wise attention and visual transformer. *Information Sciences*, 580:35–54, 2021. 2

[18] B. Jiang and K. B. Jia. Research of robust facial expression recognition under facial occlusion condition. In *International Conference on Active Media Technology*, 2011. 2

[19] S. Jiang, X. Xu, F. Liu, X. Xing, and L. Wang. Csgresnet: A simple and highly efficient network for facial expression recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2599–2603, 2022. 7

[20] A. Joshi, S. Kyal, S. Banerjee, and T. Mishra. In-the-wild drowsiness detection from facial expressions. In *2020 IEEE intelligent vehicles symposium (IV)*, pages 207–212. IEEE, 2020. 1

[21] J.-H. Kim, N. Kim, and C. S. Won. Facial expression recognition with swin transformer. *arXiv preprint arXiv:2203.13472*, 2022. 2

[22] H. Li, M. Sui, F. Zhao, Z. Zha, and F. Wu. Mvt: Mask vision transformer for facial expression recognition in the wild. *arXiv preprint arXiv:2106.04520*, 2021. 2

[23] 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*, pages 2852–2861, 2017. 7

[24] Y. Li, Y. Gao, B. Chen, Z. Zhang, G. Lu, and D. Zhang. Self-supervised exclusive-inclusive interactive learning for multi-label facial expression recognition in the wild. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):3190–3202, 2022. 1

[25] H. Liu, H. Cai, Q. Lin, X. Li, and H. Xiao. Adaptive multilayer perceptual attention network for facial expression recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):6253–6266, 2022. 7

[26] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[27] F. Ma, B. Sun, and S. Li. Robust facial expression recognition with convolutional visual transformers. *arXiv preprint arXiv:2103.16854*, 2021. 2

[28] A. Majumder, L. Behera, and V. K. Subramanian. Automatic facial expression recognition system using deep network-based data fusion. *IEEE transactions on cybernetics*, 48(1):103–114, 2016. 2

[29] N. Otberdout, M. Daoudi, A. Kacem, L. Ballihi, and S. Berretti. Dynamic facial expression generation on hilbert hypersphere with conditional wasserstein generative adversarial nets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):848–863, 2022. 1

[30] L. Pang, N. Li, L. Zhao, W. Shi, and Y. Du. Facial expression recognition based on gabor feature and neural network. In *2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, pages 489–493. IEEE, 2018. 2

[31] M. D. Putro, D.-L. Nguyen, and K.-H. Jo. A dual attention module for real-time facial expression recognition. In *IECON 2020 The 46th Annual Conference of the IEEE Industrial Electronics Society*, pages 411–416. IEEE, 2020. 2

[32] 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*, pages 7660–7669, 2021. 7

[33] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, 27(6):803–816, 2009. 1

[34] 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*, pages 6248–6257, 2021. 7

[35] J. Shi, S. Zhu, and Z. Liang. Learning to amend facial expression representation via de-albino and affinity. *arXiv preprint arXiv:2103.10189*, 2021. 7

[36] Z. Shokoohi, R. Bahmanjeh, and K. Faez. Expression recognition using directional gradient local pattern and gradient-based ternary texture patterns. In *2015 2nd International Conference on Pattern Recognition and Image Analysis (IPRIA)*, pages 1–7, 2015. 1

[37] W. Song, S. Shi, Y. Wu, and G. An. Dual-attention guided network for facial action unit detection. *IET Image Processing*, 16(8):2157–2170, 2022. 2

[38] L. Tran, X. Yin, and X. Liu. Representation learning by rotating your faces. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):3007–3021, 2019. 1

[39] 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*, pages 6897–6906, 2020. 7

[40] 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*, 29:4057–4069, 2020. 7

[41] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 2

[42] Z. Wang and Z. Ying. Facial expression recognition based on local phase quantization and sparse representation. In *2012 8th International Conference on Natural Computation*, pages 222–225, 2012. 1

[43] Z. Wen, W. Lin, T. Wang, and G. Xu. Distract your attention: Multi-head cross attention network for facial expression recognition. *arXiv preprint arXiv:2109.07270*, 2021. 7

[44] T. Wu, M. S. Bartlett, and J. R. Movellan. Facial expression recognition using gabor motion energy filters. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pages 42–47, 2010. 1

[45] M. Xia, Y. L. Xue, Z. Li, K. Huang, and S. W. Lv. Robust facial expression recognition based on rpca and adaboost. In *2009 10th Workshop on Image Analysis for Multimedia Interactive Services*, 2009. 2

[46] F. Xue, Q. Wang, and G. Guo. Transfer: Learning relation-aware facial expression representations with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3601–3610, 2021. 8

[47] F. Zhang, T. Zhang, Q. Mao, and C. Xu. A unified deep model for joint facial expression recognition, face synthesis, and face alignment. *IEEE Transactions on Image Processing*, 29:6574–6589, 2020. 1

[48] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016. 2

[49] Q. Zhang and Y.-B. Yang. Rest: An efficient transformer for visual recognition. *Advances in Neural Information Processing Systems*, 2021. 2

[50] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li. Spatial-temporal recurrent neural network for emotion recognition. *IEEE transactions on cybernetics*, pages 839–847, 2017. 2

[51] X. Zhang, F. Zhang, and C. Xu. Joint expression synthesis and representation learning for facial expression recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1681–1695, 2022. 1

[52] 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*, volume 35, pages 3510–3519, 2021. 7

[53] C. Zheng, M. Mendieta, and C. Chen. Poster: A pyramid cross-fusion transformer network for facial expression recognition. *arXiv preprint arXiv:2204.04083*, 2022. 7