The Visual Computer (2023) 39:2637–2652
https://doi.org/10.1007/s00371-022-02483-5

ORIGINAL ARTICLE

Soft thresholding squeeze-and-excitation network for pose-invariant facial expression recognition

Chaoji Liu[1] · Xingqiao Liu[1] · Chong Chen[1] · Qiankun Wang[1]

Accepted: 26 March 2022 / Published online: 26 April 2022

Abstract

Pose-invariant facial expression recognition is one of the popular research directions within the field of computer vision, but pose variant usually change the facial appearance significantly, making the recognition results unstable from different perspectives. In this paper, a novel deep learning method, namely, soft thresholding squeeze-and-excitation (ST-SE) block, was proposed to extract salient features of different channels for pose-invariant FER. For the purpose of adapting to different pose-invariant facial images better, global average pooling (GAP) operation was adopted to compute the average value of each channel of the feature map. To enhance the representational power of the network, Squeeze-and-Excitation (SE) block was embedded into the nonlinear transformation layer to filter out the redundant feature information. To further shrink the significant features, the absolute values of GAP and SE were multiplied to calculate the threshold suitable for the current view. And the developed ST-SE block was inserted into ResNet50 for the evaluation of recognition performance. In this study, extensive experiments on four pose-invariant datasets were carried out, i.e., BU-3DFE, Multi-PIE, Pose-RAF-DB and Pose-AffectNet, and the influences of different environments, poses and intensities on expression recognition were specifically analyzed. The experimental results demonstrate the feasibility and effectiveness of our method.

Keywords Pose-invariant facial expression recognition · Squeeze-and-excitation (SE) block · Soft thresholding SE block · Deep residual networks

1 Introduction

Facial expression, as the most intuitive signal for human to convey social information, has become a research hotspot in the field of human–computer interaction (HCI). Both physical and inner thoughts can be obtained through the analysis of expression variation. In previous research, various approaches have been proposed to solve the issues of facial expression recognition (FER) [1–4]. However, most of the exiting works focus on the recognition of frontal or near-frontal facial expressions, with relatively few studies on pose-variant. Nevertheless, in real-world scenarios, the captured facial images are usually determined by the angular position of the camera, which leads to rather unstable recognition accuracy [5–8]. Therefore, how to effectively extract the features based on pose-invariant images is a very challenging and meaningful task.

In the past few decades, several effective feature extraction techniques have been proposed for pose-invariant expression recognition. According to the research route, those techniques can be roughly classified into traditional based as well as deep learning-based methods. When using traditional-based methods, facial images are usually represented by geometric feature models or cropped into different regions of interest (ROIs). For example, Zhang et al. [9] used the preview-trained Active Appearance Models (AAMs) to extract the positions of facial points, and then trained each set of feature points through a specific model for pose-invariant FER. Zheng et al. [10] utilized 83 landmark points and their surrounding regions to represent facial expressions in different poses, and then extracted SIFT features for expression classification. In [11], they divided the multi-view facial images into a set of sub-blocks with the same size, and extracted LBP features from each block for FER afterward. Similarly, Zhang et al. [12] firstly presented a spatially coherent feature learning method for pose-invariant FER (SC-PFER), which

✉ Xingqiao Liu
1719618835@qq.com

[1] College of Electrical and Information Engineering, Jiangsu University, Zhenjiang City, China

normalized the expressions and poses with same horizontal and pitch angles, subsequently extracted a sequence of key regions for unsupervised feature learning, and finally used the extracted regions for FER. All these above-mentioned methods can achieve good results, but in practical application, pre-processing is an indispensable operation before feature extraction.

When using deep learning-based methods, to extract the regions of interest more accurately, numerous researches attempt to use multi-channel and multi-model feature learning methods to improve the representation ability of CNNs. As shown in Fig. 1, Liu et al. [13] presented a multi-channel pose-aware convolution neural network (MPCNN) for multi-view FER, in which channel-M1, channel-M2 and channel-M3 are used to extract whole facial region, eyes region and mouth region, respectively, and then these regions are provided to the classifier for expression recognition. Similarly, Liu et al. [14] designed a multi-channel convolution network for pose-invariant FER. The features extraction part includes three sub-CNNs, which learn different regions of interest (ROIs) of expressions, and these fusion features are fed into pose-specific CNN operations to enhance high-level feature representation. Liu et al. [15] designed a multi-channel network with pose-invariant FER, in which DML-Net is composed of three parallel channel networks, learning global and local features from different facial regions, and then integrating them for FER. It is worth mentioning that the accuracy of KDFE, BU-3DFE and Multi-PIE database are 88.2%, 83.5% and 93.5%, respectively. Moreover, in [16], they used two different channels to extract images features, and employed fixed loss weighting parameters to enhance the accuracy of expression recognition. Based on this method, Zheng et al. [17] added adaptive dynamic weight (ADW) in different channels to filter useful information, which not only reduced the chance of over-fitting, but also improved the training efficiency of the network.

Although the traditional-based and deep learning-based methods both have performed well in reducing the influence of occlusion and pose-invariant, there still remain several inevitable shortcomings. In the traditional-based setting, these methods generally require to manually crop out a large number of ROIs, which destroys the construction of automatic expression recognition system, especially the geometric feature models that are more dependent on the precise localization of feature points will be greatly limited the capabilities of following feature extraction and representation. In the deep learning-based methods setting, the multi-channel multi-model features learning methods need to not only consider the features of each region, but also pay attention to the impact of the loss function of each region on the accuracy of expression recognition, which usually results in a the convolution neural network being more complex than the traditional end-to-end networks. Moreover, using ROIs to represent the facial images in sparse pattern may not be possible to represent the original meaning of expressions completely and precisely.

In this paper, as far as the overall information of the pose-invariant expression images is concerned, the cropping of ROIs and calibration of geometric feature points is avoidable, and good operation of the automatic expression recognition system can be well ensured. All these benefits are brought by the Squeeze-and-Excitation (SE) block [20], which can dynamically recalibrate channel-wise feature in each convolutional layer in spite of the different feature maps contained in each convolutional layer, aiming to enhance the representation ability of networks on the useful layer and suppress the role of the useless layer. According to this technique, Ma et al. [18] proposed an optimized neural network based on ResNet18 and SE blocks for FER, and embedded SE model into ResNet model, which not only reduced the calculation parameters, but also improved the flow capacity of the network layer by layer. Li et al. [19] presented a Slide-Patch and Whole-Face Attention model with SE blocks (SPWFA-SE) for multi-view FER in wild condition, in which SE blocks are used as attention modules to train the weights of pre-trained patches of each channel, which can further filter out salient features from multi-view facial images. Inspired by [18, 19], in order to accommodate the different visual images, this paper proposed a soft thresholding multi-channel squeeze-and-extraction (ST-SE) block for pose-invariant FER. In each ST-SE block, the extracted feature maps were flattened by global average pooling (GAP), which were then sent into SE module. Consequently, the threshold parameters were obtained by multiplying the SE training parameters and the absolute value GAP, which could be regarded as a specific self-attention function aiming at filtering the salient features in the current views. The main contributions of this paper are summarized as follows:

1. The soft thresholding SE (ST-SE) block for pose-invariant FER is designed. Not only the SE method, but also the global average pooling (GAP) layer is added to ST-SE block. GAP operation can provide a large number of the average values from each channel of the feature map, which can force the network to pay more attention to the features in the current view.
2. The SE operation multiplied by the absolute value GAP is regarded as a self-attention mechanism, which can not only extract salient feature information, but also reduce the influence of pose-variant on the recognition accuracy.
3. In order to illustrate the effectiveness of designed ST-SE block, ResNet50 is used as the backbone architecture, as well as SE and ST-SE blocks are embedded into deep architecture as nonlinear transformation layers, respectively.

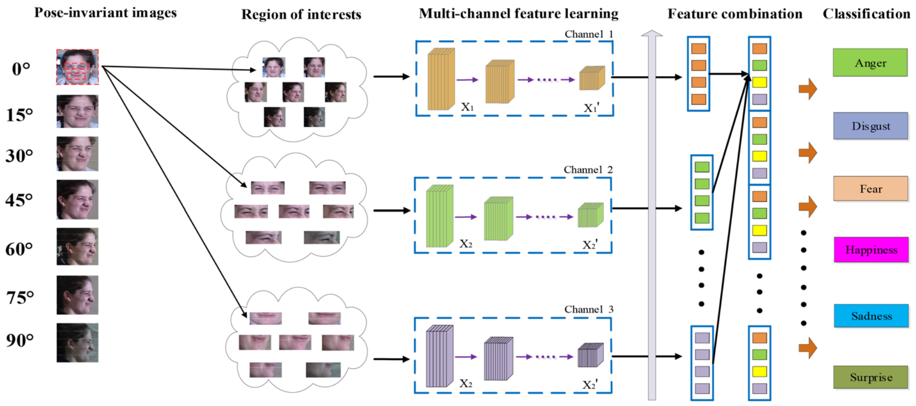


Fig. 1 The main steps of multi-channel facial expression recognition

4. This study implements extensive experiments on four public pose-invariant datasets. As shown in Fig. 2, there are not only controlled but also real-world scenario dataset, i.e., BU-3DFE, Multi-PIE, Pose-RAF-DB and Pose-AffectNet. In addition, the performance of SE and ST-SE block with some previous pose-invariant FER methods is compared, and the experiments show that the ST-SE block designed in this paper is superior.

The remainder chapters are introduced as follow: Sect. 2 introduces the related works of pose-invariant FER. Section 3 represents the proposed method in detail. The experimental results and analysis are introduced in Sect. 4. Finally, the conclusions are given in Sect. 5.

2 Related work

The ResNets and ST-SE block both contain some similar basic components, including convolutional layer, batch normalization and rectifier linear unit, which are generally considered as the essential components of convolution operations. In addition, the Global average pooling (GAP), which fully-connected layer and cross-entropy as indispensable ancillary operations, are usually utilized in deep learning to improve classification tasks. Next, this paper introduces the concepts of these components.

2.1 Basic components

Convolution layer (Conv) is a role component that implements the convolution operation to the input image for extracting feature maps and then transmits them to the next layer. Each convolutional layer consists of a plurality of neurons with trainable weight and biases, and each feature map is implemented by a convolutional kernel over the input channels with fixed stride, which can be defined as follows:

$$x_j^{l+1} = \sum_{i \in M_j} x_i^l * k_{ij}^l + b_j^l \tag{1}$$

where x_i^l denotes the input feature map at the ith channel, x_i^{l+1} denotes the output feature map at the jth channel, k denotes the weight matrix of the convolutional kernel, b denotes the bias, and M_j denotes one of the feature maps in convolution layer.

As a feature normalizing method, batch normalization (BN) is usually inserted by convolution layer to accelerate the convergence of network training [21]. The BN plays a role in decreasing the offset of internal covariates during the process of training deep learning network. Especially in pose-invariant context, the distribution of training data usually varies with different views. BN operation can normalize the features of activation values to a fixed distribution during the training process, and adjust the feature mapping within a reasonable distribution range, which is an essential operation in a very deep network. The calculation steps can be expressed as follows:

$$\mu_{(N_{batch})} = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{2}$$

$$\sigma^2_{(N_{batch})} = \frac{1}{m}\sum_{i=1}^{m} \left(x_i - \mu_{(N_{batch})}\right)^2 \tag{3}$$

$$\hat{x}_i = \frac{x_i - \mu_{(N_{batch})}}{\sqrt{\sigma^2_{N_{batch}} + \varepsilon}} \tag{4}$$

$$y_i = \gamma \hat{x}_i + \beta \tag{5}$$

where x_i and y_i denote the input and output feature maps in current batch, respectively. m denotes the batch size. γ and β denote scale factor and movement factor, respectively. ε denotes a constant, which is composed to avoid meeting undefined of $\sqrt{s}$ at $s = 0$.

The rectifier linear unit (ReLU) serves as the other indispensable component of convolution operations whose appears and behaves are similar to a linear function, but instead are non-saturated and nonlinear features enabling complex layers of input feature maps to be learned. For any positive input x, the output is the same value. However, while

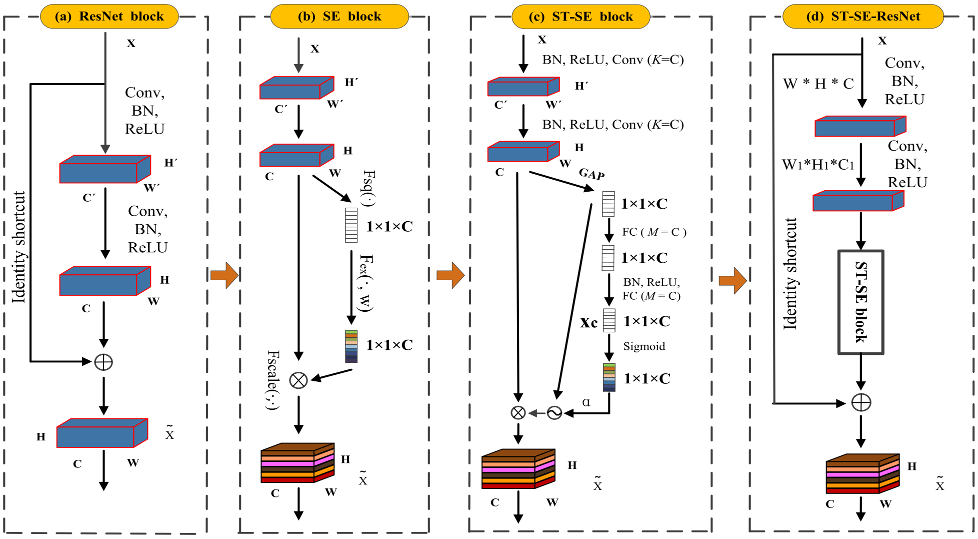


Fig. 2 An over view of the proposed ST-SE-ResNet block. **a** The ResNet block is used to extract feature map. **b** SE block is used to extract prominent features from different channels. **c** The ST-SE block is a soft thresholding operation that forces prominent features in current layer. **d** A basic ST-SE-ResNet block. $\tilde{x}$ and α denote the candidate feature maps when the threshold is determined

the input is negative, the input will be forced to be 0, which can be expressed as $f(x) = \max(0, x)$, and can cleverly solve the problems of gradient vanishing and gradient exploding when the parameters are trained among different layers.

2.2 Global average pooling

Global average pooling (GAP) is another indispensable operation that computes the average value from each channel of the feature map [22]. Similar to fully connected (FC) layers, it is usually applied for the last layer in the entire conventional structure. However, since there no parameters to be optimized, GAP can use less weights compared to FC layer, which reduces the possibility of overfitting. In addition, it needs to be mentioned that GAP can also solve the shift variant problem, which provides a unique advantage for pose-invariant and complex environmental background.

2.3 Fully connection layer

The fully connected (FC) layer is similar to the multi-perceptron neural net-works, and the neuron activation is fully connected with previous layer. The number of neuron activation in the last layer is determined by the input convolution kernel, and FC operation can flatten the input into a single vector in the next layer. Therefore, FC layer contains a large amount of parameters that characterize the characteristics and laws of sample data. For some classical convolutional models, i.e., VGG, GoogLeNet and ResNet, 1–3 FCs can generally solve the complex image classification problems.

2.4 Loss function

With respect to the loss function, cross-entropy is one of the most well-known loss functions in FER tasks. Before implementing cross-entropy operation, a softmax function is usually executed to limit the features range within (0, 1). It can be defined as follows:

$$y_j = \frac{e^{x_j}}{\sum_{i=1}^{N_{class}} e^{x_i}} \tag{6}$$

where x_j denotes the jth input feature map of softmax function, y_j denotes a predicted probability belong to jth class, N_{class} denotes the number of classes. Then the cross-entropy loss function can be expressed as:

$$E(p(y), q(y)) = -\sum_{j=1}^{N_{class}} p_j(y) \log(q_j(y)) \tag{7}$$

where $p(x)$ denotes the target values, $q(x)$ denotes the real probability of x belonging to the jth class.

3 Proposed method

From the above, it can be seen that both residual network and SE block are composed of these basic elements. In this section, this paper presents in detail the improvement process of soft thresholding multi-channel SE Residual Network structure (ST-SE-ResNet). As shown in Fig. 2, this study first introduces the residual network, then describes the SE block, next describes the ST-SE, and finally introduces ST-SE-ResNet block.

3.1 Residual building blocks

ReseNet is a classical network model with “identity shortcut layer”, which has been widely concerned by scientific researchers [23]. As shown in Fig. 2a, the basic residual block (RBBs) consists of two BNs, two ReLUs, two Conv-layers and an identity shortcut layer. The key operation is identity shortcut that effectively back-propagates the gradient of loss function to earlier layers, which makes ResNet superior to the traditional deep learning methods. The residual block is described as:

$$F(X_{re\sin}) = H(X_{re\sin}) - X_{re\sin} \tag{8}$$

where $X_{re\sin}$ denotes the input feature map, $H(X_{re\sin})$ denotes the desired feature maps and $F(X_{re\sin})$ denotes the output feature maps of one residual module.

3.2 Squeeze-and-excitation block

As mentioned in [18–20], a multi-channel SE block was implemented to improve the representation of feature. The function of SE block is to learn feature information from different channels that can enhance the representation ability by a single basic block. As shown in Fig. 2b, for each input channel, a weight can be trained by a basic SE block. Here we assume $X = \{x_1, x_2, \cdots x_n\}$ is the input feature map of SE block and Z_c=$\{z_1, z_2, \cdots z_n\}$ is the corresponding output feature map, the Squeeze operation is described as:

$$z_c = F_{sq}(x_c) = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} x_c(i, j), c = 1, 2, \cdots, n \tag{9}$$

where W and H denote the width and height of input feature maps of SE block, z_c denote the output of current layer. n denote the channel in SE block.

To enhance the representation ability from the current convolutional layer, Excitation operation is described as:

$$s_c = F_{ex}(z_c, \omega) = \sigma(f(z_c, \omega)) = \sigma(\omega_2 \delta(\omega_1 z_c)) \tag{10}$$

where ω_1 and ω_2 denote the weight matrices in two FC layers. δ and σ denote ReLU and sigmoid function.

$$\tilde{x}_c = F_{scale}(x_c, s_c) = s_c x_c \tag{11}$$

where $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_c)$ and $F_{scale}(x_c, s_c)$ denote channel-wise multiplication between scaling parameter s_c and feature map $x_c \in \mathbb{R}^{H \times W}$.

3.3 Soft thresholding SE block

The designed soft thresholding SE block (ST-SE) is a variant of SE, and its main difference is that a specific threshold is learned by each channel of the feature map, meaning that each channel can learn a specializing threshold to refine the significant feature information under current layer of FER. As shown in Fig. 2c, the ST-SE-ResNet block contains a special model, where GAP is used to flatten the feature map into a 1D vector. Next, the 1D vector is sent to two fully-connected layers to obtain a training parameter, the operation of which is similar to the SE block [20], and the number of convolutional cores is equal to the numbers of channels. Finally, the sigmoid function is used to keep the training parameters within the range of (0, 1), and the operation is described as follows:

$$\alpha_c = \frac{1}{1 + e^{-x_c}} \tag{12}$$

where x_c denotes the output two fully connected layers, and α_c denotes the cth training parameter. Next, the training parameter α and $|x|$ are multiplied to obtain the threshold. The inspiration of this design is the fact that the threshold parameters need to be positive and not too large. Owing to a pose-invariant FER setting, the views have a very obvious influence on the recognition accuracy, especially on the edge. In order to reduce the impact of posture and background, the threshold values in a ST-SE-ResNet block are calculated as follows:

$$\tau_c = \alpha_c \cdot \underset{i,j}{average} \left| x_{i,j,c} \right| \tag{13}$$

where τ_c denotes the threshold in the cth channel, i, j, and c denotes the index of width, height and channel of feature map x, respectively.

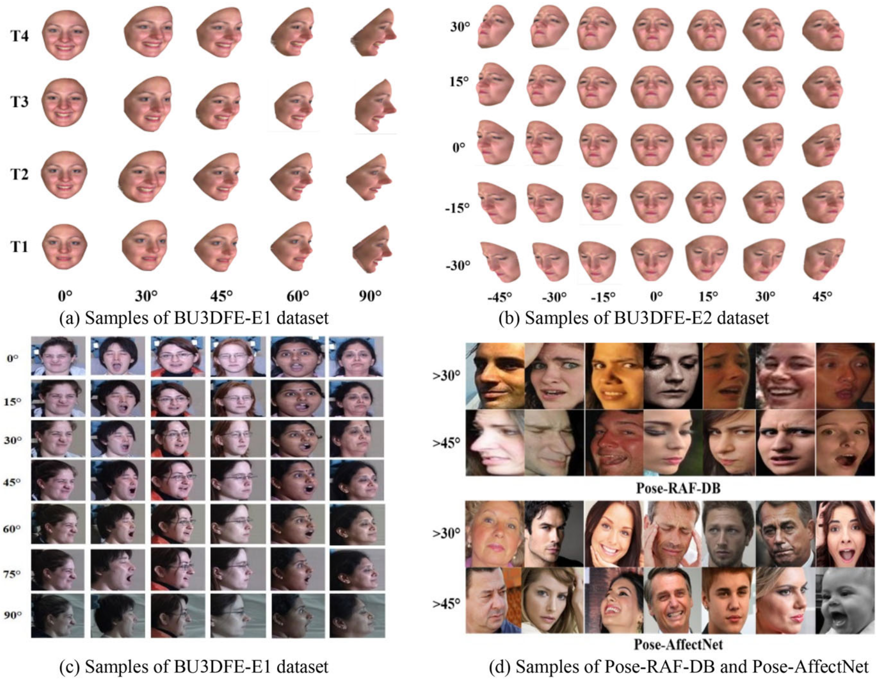
To demonstrate the practical use of the proposed ST-SE module, it is vital to construct the same network structure and parameter settings. Considering the diversity of the expression images in the same view pictures, this paper uses ResNet50 as the basic network architecture, and embeds the SE and ST-SE modules in the network, respectively, as shown in Fig. 2d. The architectures of ResNet50, SE-ResNet50 and ST-SE-ResNet50 are listed in Table 1.

4 Experimental results

To evaluate the effectiveness of the designed network, this paper performed extensive experiments on four famous facial expression databases that are BU-3DFE [24] and Multi-PIE [25] which were collected in a controlled environment, as

Table 1 The parameters of ResNet50 (Left), SE-ResNet50 (Middle), ST-SE-ResNet50 (Right), *fc* denotes two fully connected layers in a SE-ResNet50 basic block

Output size	Output size	ResNet50	SE-ResNet50	ST-SE-ResNet50
Conv1_x	112 × 112	Conv, 7 × 7, 64, stride 2 Max pool, 3 × 3, stride 2	Conv, 7 × 7, 64, stride 2 Max pool, 3 × 3, stride 2	Conv, 7 × 7, 64, stride 2 Max pool, 3 × 3, stride 2
Conv2_x	56 × 56	$\begin{bmatrix} conv, & 1\times1, & 64 \\ conv, & 3\times3, & 64 \\ conv, & 1\times1, & 256 \end{bmatrix}\times 3$	$\begin{bmatrix} conv, & 1\times1, & 64 \\ conv, & 3\times3, & 64 \\ conv, & 1\times1, & 256 \\ fc, & [16, 256] & \end{bmatrix}\times 3$	$\begin{bmatrix} conv, & 1\times1, & 64 \\ conv, & 3\times3, & 64 \\ conv, & 1\times1, & 256 \\ fc, & [16, 256] & \end{bmatrix}\times 3$
Conv3_x	28 × 28	$\begin{bmatrix} conv, & 1\times1, & 128 \\ conv, & 3\times3, & 128 \\ conv, & 1\times1, & 512 \end{bmatrix}\times 4$	$\begin{bmatrix} conv, & 1\times1, & 128 \\ conv, & 3\times3, & 128 \\ conv, & 1\times1, & 512 \\ fc, & [32, 512] & \end{bmatrix}\times 4$	$\begin{bmatrix} conv, & 1\times1, & 128 \\ conv, & 3\times3, & 128 \\ conv, & 1\times1, & 512 \\ fc, & [32, 512] & \end{bmatrix}\times 4$
Conv4_x	14 × 14	$\begin{bmatrix} conv, & 1\times1, & 256 \\ conv, & 3\times3, & 256 \\ conv, & 1\times1, & 1024 \end{bmatrix}\times 6$	$\begin{bmatrix} conv, & 1\times1, & 256 \\ conv, & 3\times3, & 256 \\ conv, & 1\times1, & 1024 \\ fc, & [64, 1024] & \end{bmatrix}\times 6$	$\begin{bmatrix} conv, & 1\times1, & 256 \\ conv, & 3\times3, & 256 \\ conv, & 1\times1, & 1024 \\ fc, & [64, 1024] & \end{bmatrix}\times 6$
Conv5_x	7 × 7	$\begin{bmatrix} conv, & 1\times1, & 512 \\ conv, & 3\times3, & 512 \\ conv, & 1\times1, & 2048 \end{bmatrix}\times 3$	$\begin{bmatrix} conv, & 1\times1, & 512 \\ conv, & 3\times3, & 512 \\ conv, & 1\times1, & 2048 \\ fc, & [128, 2048] & \end{bmatrix}\times 3$	$\begin{bmatrix} conv, & 1\times1, & 512 \\ conv, & 3\times3, & 512 \\ conv, & 1\times1, & 2048 \\ fc, & [128, 2048] & \end{bmatrix}\times 3$
	1 × 1	Average pool, 1000-d *fc*, softmax	Average pool, 1000-d *fc*, softmax	Average pool, 1000-d *fc*, softmax



(a) Samples of BU3DFE-E1 dataset

(b) Samples of BU3DFE-E2 dataset

(c) Samples of BU3DFE-E1 dataset

(d) Samples of Pose-RAF-DB and Pose-AffectNet

Fig. 3 Some examples of the two datasets (BU3DFE-E1, BU3DFE-E2, Multi-PIE, Pose-RAF-DB and Pose-AffectNet)

well as RAF-DB [26] and AffectNet [27] captured in real-world scenarios. Some samples of these databases are shown in Fig. 3. Since the BU-3DFE and Multi-PIE databases did not precisely divide training and testing sets, in this work, the fivefold cross-validation protocol was employed in these databases. The designed ST-SE-ResNet50 framework was carried out on PyTorch, and the learning rate and the batch size were set to 0.000001 and 40, respectively. The size of the input images was adjusted to 224 × 224, because using large images could improve the deep learning ability of the network, in which more salient features were able to be extracted. All the experiments was based on NVIDIA GeForce GTX 1660 Super GPU; Operating System: Windows 10 64bits.

4.1 Experiments with BU-3DFE dataset

This paper first tested the designed network on BU-3DFE dataset, which was widely used in pose-invariant FER. There were a total of 100 subjects involved in the experiment, and each of them contained 6 typical expressions, i.e., anger (AN), disgust (DI), happiness (HA), fear (FE), sadness (SA) and surprise (SU) in 4 different intensities. Before using the original dataset, the 3D expression models were typically rotated on the invariant views to generate 2D texture images. Among the existing pose-invariant FER methods, two mainstream methods of extended 2D facial expression image sets were widely adopted. Next, this paper performed experiments on these two extended pose-invariant datasets, and compared the results with some previous methods.

For the first extended dataset of BU-3DFE (BU3DFE-E1), it contains 5 × 4 × 6 × 100 = 12000 2D texture expression images in 5 invariant yaw angles (0°, 30°, 45°, 60°, 90°) from 4 different intensities. The corresponding expression images are shown in Fig. 3a, many previous works [10, 11, 30–32], adapted BU3DFE-E1 dataset for pose-invariant FER experiments and achieved remarkable results. This paper evaluated these three network framework structures on the BU3DFE-E1 dataset and analyzed the reasons for these results.

As show in Table 2, this study compares the results of ST-SE-ResNet50 method with SE-ResNet50, ResNet50 and some previous works on BU3DFE-E1 dataset. It is worth mentioning that BU3DFE-E1 dataset contains not only 5 invariant yaw angles, but also 4 different intensities. The SE-ResNet50 network achieved 75.9% recognition accuracy, which was a little better than that of basic ResNet50. In contrast, the ST-SE-ResNet50 model could further improve the identification accuracy to 76.20%. Especially, the pose-invariant recognition algorithms that was often referenced was superior to 2D JFDNN (72.5%), CNN (68.9%), VGGNet16 (70.1%), and slightly better than the DBN (73.5%) and LLCBL (74.60%). For the classical Local binary patterns (LBP), the designed network was 5.1% higher than the highest recognition accuracy.

Table 3 lists the specific recognition accuracy of 6 typical expressions under five yaw angles, and Fig. 4 shows the corresponding confusion matrix. In Table 3, it easy to find that the recognition accuracy varies with the yaw angle, where the best yaw angle of expression recognition is 60° with the accuracy rate of 78.6%, while the worst yaw angle is 90° with the accuracy rate of 73.6%. In addition, for the 6 basic typical expressions, the performances of recognition accuracies are also different. Happiness and surprise as the most obvious expressions to distinguish are usually the easiest to recognize in all different yaw angles, while fear is the most challenging expressions, whose recognition accuracies are less than 63%. Figure 4 shows the expression confusion matrix in each yaw view, we can see that angry and sadness expressions are more easily confused, which is the reason why the recognition accuracies of these two expressions are low. In the meantime, on the whole, all the misclassification rates of fear expression relatively higher than other expressions, which lead to the lowest recognition accuracy of fear among the six typical expressions.

The second extended dataset of BU-3DFE (BU3DFE-E2) contains 7 × 5 × 6 × 100=21000 2D texture expression images with 7 invariant pan angles (0°, ±15°, ±30°, ±45°) and 5 invariant tilt angles (0°, ±15°, ±30°). The corresponding expression images are shown in Fig. 3b. Compared with BU3DFE-E1 dataset, BU3DFE-E2 pays more attention on the impact of different views on expression recognition. For example, the BU3DFE-E1 dataset only contains a pan angle, however, the pan angles are extended from − 45° to + 45° and the title angles are set vary from − 30° to + 30° in BU3DFE-E2 dataset. Besides, the BU3DFE-E2 dataset comprises only the 4th intensity level of 2D texture expression images, but the images in the BU3DFE-E1 dataset contains all intensity levels. Some of the state-of-the-art methods [34, 35] also adopt BU3DFE-E2 dataset for pose-invariant FER experiments. This paper evaluates the proposed method with all these expression images at 7 invariant yaw angles.

In the same way as BU3DFE-E1 dataset, this paper compares the method with previous works [9, 30, 32, 33] and presents the results in Table 4. It can be seen that ST-SE-ResNet50 achieves 83.7%, while the best result among state-of-the-art method is only 81.2%, which is far lower than the algorithm in this paper. Moreover, the recognition accuracy of ST-SE-ResNet50 is 4.1% higher than that of basic ResNet50, which demonstrates that the designed method also performs well under mixed multi-view. Table 5 lists the specific recognition accuracy rates under different angles, where the best yaw angle of expression classification is −30° with the accuracy rate of 85.17%, and the worst yaw angle is 45° with the accuracy rate of 82.83%. In addition, for the average recognition results of each expression, they are roughly

Table 2 The comparison with different methods on the BU3DFE-E1 dataset

Method	Pose			Expressions		Feature	Accuracy (%)
	Number	Pan	Tilt	Number	Levels		
Zheng et al. [10]	5	(0°, + 90°)	(−0°, + 0°)	6	1, 2, 3, 4	LBP	66.0
Moore et al. [11]	5	(0°, + 90°)	(−0°, + 0°)	6	1, 2, 3, 4	LBP	65.0
Moore et al. [11]	5	(0°, + 90°)	(−0°, + 0°)	6	1, 2, 3, 4	LGBP	68.0
Moore et al. [11]	5	(0°, + 90°)	(−0°, + 0°)	6	1, 2, 3, 4	LGBP/LBP	71.1
Zhang et al. [30]	5	(0°, + 90°)	(−0°, + 0°)	6	1, 2, 3, 4	CNN	68.9
Zhang et al. [31]	5	(0°, + 90°)	(−0°, + 0°)	6	1, 2, 3, 4	DBN	73.5
Wu et al. [28]	5	(0°, + 90°)	(−0°, + 0°)	6	1, 2, 3, 4	LLCBL	74.6
Jung et al. [29]	5	(0°, + 90°)	(−0°, + 0°)	6	1, 2, 3, 4	2D JFDNN	72.5
Zhang et al. [12]	5	(0°, + 90°)	(−0°, + 0°)	6	1, 2, 3, 4	VGGNet16	70.1
ResNet50	5	(0°, + 90°)	(−0°, + 0°)	6	1, 2, 3, 4	ResNet50	**74.8**
SE-ResNet50(**Ours**)	5	(0°, + 90°)	(−0°, + 0°)	6	1, 2, 3, 4	SE	**75.9**
ST-SE-ResNet50(**Ours**)	5	(0°, + 90°)	(−0°, + 0°)	6	1, 2, 3, 4	ST-SE	**76.2**

Table 3 Average recognition accuracies under different yaw angles on BU3DFE-E1 dataset

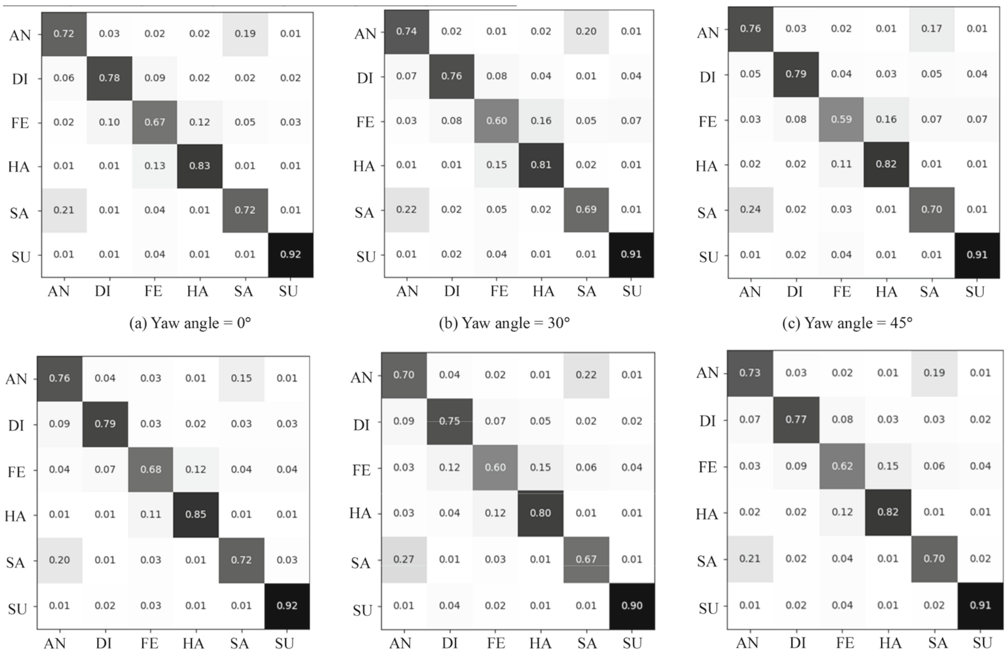
Expression	Results (%)					
	0°	30°	45°	60°	90°	Average
Angry	72.0	74.0	76.0	76.0	70.0	73.6
Disgust	78.0	76.0	**79.0**	79.0	75.0	77.4
Fear	67.0	60.0	59.0	68.0	60.0	62.8
Happy	83.0	81.0	82.0	**85.0**	80.0	82.2
Sad	72.0	69.0	70.0	71.0	67.0	70.0
Surprise	**92.0**	91.0	91.0	**92.0**	90.0	91.2
Average	77.3	75.1	76.1	**78.6**	73.6	76.2

consistent with BU3DFE-E1 dataset. As shown in Fig. 5h, the fear is still the most challenging expressions, and anger and sadness are also the most confusing expressions, but the overall recognition accuracy of each expression has been significantly improved.

4.2 Experiments with multi-PIE dataset

The Multi-PIE captured by a closely real-world scene contained 755,370 facial expression images of 377 different samples. Unlike BU-3DFE database, it included not only different postures, but also unbalanced illumination, and background transformations. In this work, same experimental setting for facial expression recognition as reporting in [11, 28–30, 34] was adopted, where only 100 subjects presented in all four recording sessions were selected. For each subject, six different emotions (disgust, neural, scream, smile, squint and surprise) and 7 invariant yaw angles (0°, 15°, 30°, 45°, 60°, 75° and 90°) were selected in the experiments. Therefore, a total of 7 × 6 × 100=4200 images were selected in the experiments. The sample of six subjects performing 42 facial images can be found in Fig. 2c.

Compared with the ResNet50 and SE-ResNet50 at 7 invariant yaw angles, the corresponding average recognition accuracy are 80.0% and 83.1%, respectively, while the method proposed in this paper is 86.1%, which is higher than other methods. Table 7 lists the specific identification results under different facial yaw angles, where the optimal expression recognition is different for each angle of view, and the best ones are often kept between 0° and 30°, which are 88.1%, 87.3% and 89.0%, respectively. Figure 6 shows each yaw angle and the overall confusion matrices. It can be seen from Table 7 and Fig. 6 that among the six typical expressions, the recognition accuracies of scream and surprise are higher, and their average recognition rates are 96.4% and 92.6%, respectively. On the contrary, squint and disgust, as the most difficult expressions to identify, has the average recognition accuracies less than 80%. Moreover, from the overall confusion matrix, we can see that the expressions squint and disgust



	AN	DI	FE	HA	SA	SU
AN	0.72	0.03	0.02	0.02	0.19	0.01
DI	0.06	0.78	0.09	0.02	0.02	0.02
FE	0.02	0.10	0.67	0.12	0.05	0.03
HA	0.01	0.01	0.13	0.83	0.01	0.01
SA	0.21	0.01	0.04	0.01	0.72	0.01
SU	0.01	0.01	0.04	0.01	0.01	0.92

(a) Yaw angle = 0°

	AN	DI	FE	HA	SA	SU
AN	0.74	0.02	0.01	0.02	0.20	0.01
DI	0.07	0.76	0.08	0.04	0.01	0.04
FE	0.03	0.08	0.60	0.16	0.05	0.07
HA	0.01	0.01	0.15	0.81	0.02	0.01
SA	0.22	0.02	0.05	0.02	0.69	0.01
SU	0.01	0.02	0.04	0.01	0.01	0.91

(b) Yaw angle = 30°

	AN	DI	FE	HA	SA	SU
AN	0.76	0.03	0.02	0.01	0.17	0.01
DI	0.05	0.79	0.04	0.03	0.05	0.04
FE	0.03	0.08	0.59	0.16	0.07	0.07
HA	0.02	0.02	0.11	0.82	0.01	0.01
SA	0.24	0.02	0.03	0.01	0.70	0.01
SU	0.01	0.02	0.04	0.01	0.01	0.91

(c) Yaw angle = 45°

	AN	DI	FE	HA	SA	SU
AN	0.76	0.04	0.03	0.01	0.15	0.01
DI	0.09	0.79	0.03	0.02	0.03	0.03
FE	0.04	0.07	0.68	0.12	0.04	0.04
HA	0.01	0.01	0.11	0.85	0.01	0.01
SA	0.20	0.01	0.03	0.01	0.72	0.03
SU	0.01	0.02	0.03	0.01	0.01	0.92

(d) Yaw angle = 60°

	AN	DI	FE	HA	SA	SU
AN	0.70	0.04	0.02	0.01	0.22	0.01
DI	0.09	0.75	0.07	0.05	0.02	0.02
FE	0.03	0.12	0.60	0.15	0.06	0.04
HA	0.03	0.04	0.12	0.80	0.01	0.01
SA	0.27	0.01	0.03	0.01	0.67	0.01
SU	0.01	0.04	0.02	0.01	0.01	0.90

(e) Yaw angle = 90°

	AN	DI	FE	HA	SA	SU
AN	0.73	0.03	0.02	0.01	0.19	0.01
DI	0.07	0.77	0.08	0.03	0.03	0.02
FE	0.03	0.09	0.62	0.15	0.06	0.04
HA	0.02	0.02	0.12	0.82	0.01	0.01
SA	0.21	0.02	0.04	0.01	0.70	0.02
SU	0.01	0.02	0.04	0.01	0.02	0.91

(f) Overall

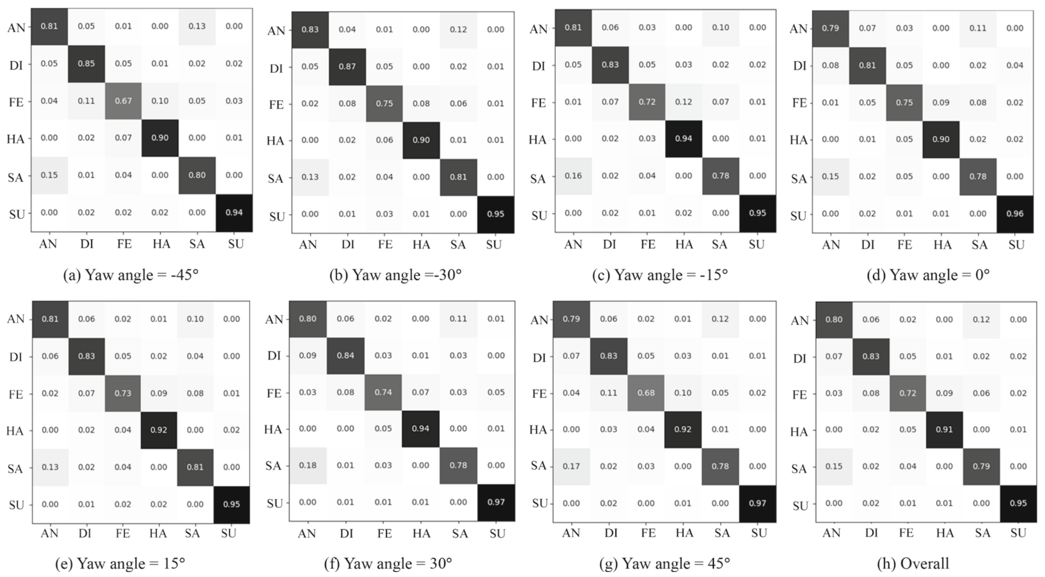
Fig.4 The confusion matrices on the BU3DFE-E1 dataset. Where **a–e** denotes the confusion matrices of five invariant yaw angle, and **f** denotes the overall recognition confusion matrices

Table 4 The comparison with different methods on the BU3DFE-E2 dataset

Method	Pose			Expressions		Feature	Accuracy (%)
	Number	Pan	Tilt	Number	Levels		
Zhang et al. [9]	35	(−45°, + 45°)	(−30°, + 30°)	6	4	Geometry features	76.6
Jampour et al. [33]	35	(−45°, + 45°)	(−30°, + 30°)	6	4	HOG + LBP	78.6
Wu et al. [30]	35	(−45°, + 45°)	(−30°, + 30°)	6	4	LLCBL	80.2
Zhang et al. [30]	35	(−45°, + 45°)	(−30°, + 30°)	6	4	DBN	75.2
Zhang et al. [31]	35	(−45°, + 45°)	(−30°, + 30°)	6	4	CNN	77.2
Zhang et al. [31]	35	(−45°, + 45°)	(−30°, + 30°)	6	4	GAN	81.2
Can et al. [32]	35	(−45°, + 45°)	(−30°, + 30°)	6	4	VGGNet16	73.1
ResNet50	35	(−45°, + 45°)	(−30°, + 30°)	6	4	ResNet50	**79.6**
SE-ResNet50	35	(−45°, + 45°)	(−30°, + 30°)	6	4	SE	**82.2**
ST-SE-ResNet50	35	(−45°, + 45°)	(−30°, + 30°)	6	4	ST-SE	**83.7**

Table 5 Average recognition accuracies under invariant angles on BU3DFE-E2 database

Expression	Results (%)							
	−45°	−30°	−15°	0°	15°	30°	45°	Average
Angry	81.0	**83.0**	81.0	79.0	81.0	80.0	79.0	80.5
Disgust	85.0	**87.0**	83.0	81.0	83.0	84.0	83.0	83.7
Fear	67.0	**75.0**	72.0	**75.0**	73.0	74.0	68.0	72.0
Happy	90.0	90.0	**94.0**	90.0	92.0	**94.0**	92.0	91.7
Sad	80.0	**81.0**	78.0	78.0	**81.0**	78.0	78.0	79.1
Surprise	94.0	95.0	95.0	96.0	95.0	**97.0**	**97.0**	95.5
Average	82.8	**85.1**	83.8	83.1	84.1	84.5	82.8	83.5



(a) Yaw angle = -45°

(b) Yaw angle =-30°

(c) Yaw angle = -15°

(d) Yaw angle = 0°

(e) Yaw angle = 15°

(f) Yaw angle = 30°

(g) Yaw angle = 45°

(h) Overall

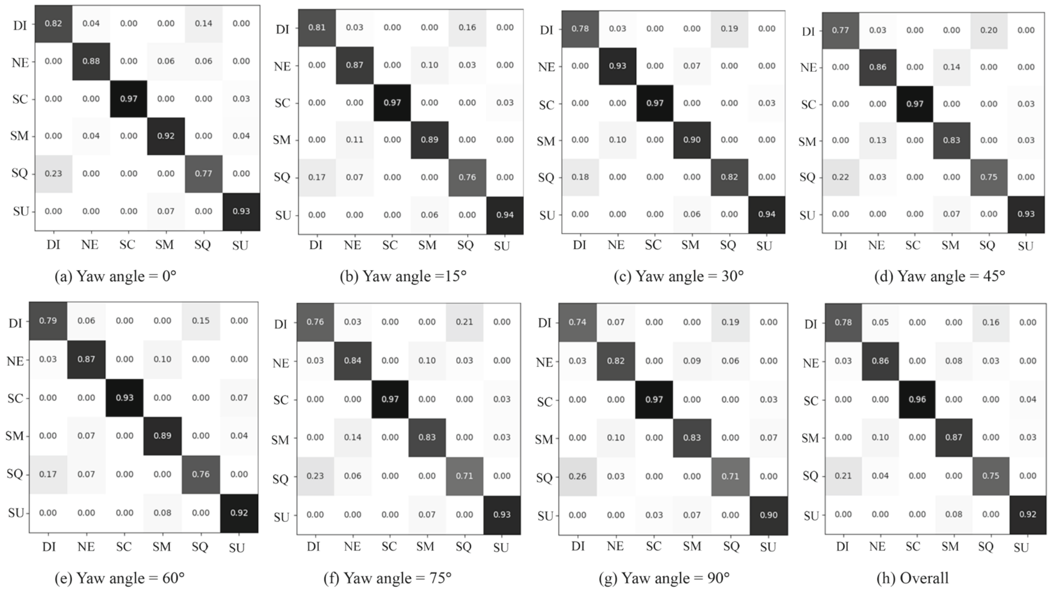
Fig.5 The confusion matrices on the BU3DFE-E2 database. Where **a**–**g** denotes the confusion matrices of 7 invariant yaw angle, and **h** denotes the overall recognition confusion matrices

Table 6 The comparison with different methods on the Multi-PIE dataset

Method	Poses		Expressions number	Feature	Accuracy (%)
	Number	Pan			
Moore et al. [11]	7	(0°, + 90°)	6	LBP	73.3
Moore et al. [11]	7	(0°, + 90°)	6	LGBP	80.4
Zhang et al. [30]	7	(0°, + 90°)	6	DBN	76.1
Zhang et al. [30]	7	(0°, + 90°)	6	CNN	77.8
Jung et al. [29]	7	(0°, + 90°)	6	JFDNN	82.9
Wu et al. [28]	7	(0°, + 90°)	6	LLCBL	80.9
Fan et al. [34]	7	(0°, + 90°)	6	VGG16	71.7
ResNet18	7	(0°, + 90°)	6	ResNet18	80.13
ResNet50	7	(0°, + 90°)	6	ResNet50	**81.0**
SE-ResNet50	7	(0°, + 90°)	6	SE	**83.1**
ST-SE-ResNet50	7	(0°, + 90°)	6	ST-SE	**86.1**

Table 7 Average recognition accuracies under invariant angles on Multi-PIE dataset

Expressions	Results (%)							Average
	0°	15°	30°	45°	60°	75°	90°	
Disgust	**82.0**	81.0	78.0	77.0	79.0	76.0	74.0	78.1
Neutral	88.0	87.0	**93.0**	86.0	87.0	84.0	82.0	86.7
Scream	97.0	97.0	**97.0**	97.0	93.0	97.0	97.0	96.4
Smile	**92.0**	89.0	90.0	83.0	89.0	83.0	83.0	87.0
Squint	77.0	76.0	**82.0**	75.0	76.0	71.0	71.0	75.4
Surprise	93.0	94.0	**94.0**	93.0	92.0	93.0	90.0	92.6
Average	88.1	87.3	89.0	85.1	86.0	84.8	82.8	86.1



(a) Yaw angle = 0°

(b) Yaw angle =15°

(c) Yaw angle = 30°

(d) Yaw angle = 45°

(e) Yaw angle = 60°

(f) Yaw angle = 75°

(g) Yaw angle = 90°

(h) Overall

Fig.6 The confusion matrices on the Multi-PIE dataset. Where **a**-**g** denotes the confusion matrices of 7 invariant yaw angle, and (h) denotes the overall recognition confusion matrices

are more prone to misclassification, and the most likely the fact that they achieve low recognition accuracies.

4.3 Experiments with pose-RAF-DB and pose-AffectNet dataset

To evaluate the performance of the model in pose-invariance under real-world scenarios, Wang [35] et al. also collected two sub-datasets, namely, Pose-RAF-DB and Pose-AffectNet, respectively, from the test datasets of FAF-DB and AffectNet for facial expression recognition. Where the head pitch or yaw angles larger than 30° and 45° were selected as a set of pose-invariant facial images, and 7 typical expressions (anger (AN), disgust (DI), happiness (HA), fear (FE), neutral (NE), sadness (SA) and surprise (SU)) were considered. Therefore, Pose-RAF-DB consisted of 12,271 facial images for training, and 1,248 and 558 facial images were selected to test sub-datasets at 30° and 45°, respectively, while Pose-AffectNet consisted of 283,901 facial images for training, and 1,948 and 985 facial images were selected to test sub-datasets at 30° and 45°, respectively. In this work, the same experiment setting in [35, 35, 37, 38] was adopted. However, it is needed to mention that both Pose-RAF-DB and Pose-AffectNet treat forward and reverse facial images as a pose-invariant dataset, which enriches the database and increases the difficulty of classification, as shown in Fig. 3d.

This paper conducted experiments on databases Pose-RAF-DB and Pose-AffectNet, and the experimental results are listed in Table 8. The recognition accuracies of ST-SE-ResNet50 on Pose-FAF-DB database were 85.00% (> 30°) and 84.42% (> 45°), while those on Pose-AffectNet database were 56.57% (> 30°) and 57.00% (> 45°), respectively. Figure 7 shows the corresponding confusion matrices, from which can be found that happiness is the easiest recognizable expression in all databases; Sadness is relatively easy to identify in the Pose-FAF-DB dataset; Fear is relatively easily identified in Pose-AffectNet dataset; and disgust is the most difficult expression to classify. It is easier to be confused with neutral in Pose-RAF-DB dataset, and it is generally confused with anger in Pose-AffectNet database, which reduces the expression recognition accuracy in these two datasets.

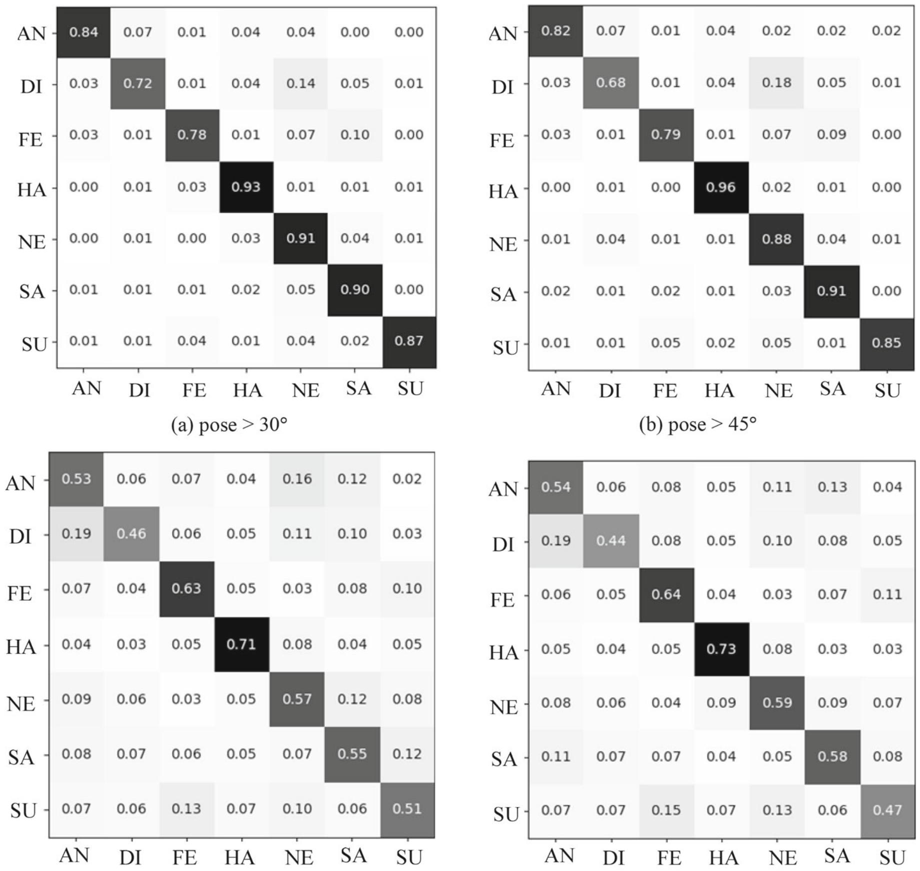
4.4 Experimental results analysis

From the recognition results of four pose-invariant datasets, it can be found that multi-channel soft thresholding SE residual network can achieve the same accuracy as the previous methods. Compared with the original Resnet50, the accuracy of the method in this paper on BU3DFE-E1, BU3DFE-E2, Multi-PIE, Pose-RAF-DB and Pose-AffectNet dataset can further improve by 1.6%, 4.1% and 5.1%, (0.44% (> 30°), 0.15% (> 45°)) and (0.26% (> 30°), 0.08% (> 45°)), respectively. Two reasons can explain this improvement. The first is that squeeze-and-excitation (SE) block serves as a bridge between different channels, whose function is to improve the quality of the designed network by using the interdependencies between the channels of its convolutional features; And the second is that GAP and SE block can be regarded as an attention mechanism, whose task is to learn global information to selectively emphasize valuable features from the

Table 8 Average recognition accuracies under invariant angles on Pose-RAF-DB and Pose-AffectNet dataset

Method	Pose		Pose-RAF-DB		Pose-AffectNet	
	Number	Pan	Pose > 30	Pose > 45	Pose > 30	Pose > 45
Wang et al. [35]	7	± (30°, 90°)	86.74	85.20	53.90	53.19
Gera et al. [36]	7	± (30°, 90°)	86.12	84.41	59.17	57.66
Gera et al. [37]	7	± (30°, 90°)	89.82	89.07	60.41	60.86
Zhao et al. [38]	7	± (30°, 90°)	87.89	87.99	57.51	57.78
VGG16	7	± (30°, 90°)	81.27	80.15	51.94	52.33
ResNet18	7	± (30°, 90°)	84.04	83.15	56.31	56.62
ResNet50	7	± (30°, 90°)	84.56	84.27	56.38	56.83
SE-ResNet50	7	± (30°, 90°)	86.67	87.17	58.82	58.79
ST-SE-ResNet50	7	± (30°, 90°)	**85.00**	**84.42**	**56.57**	**57.00**

Fig.7 Where **a**-**b** denotes the confusion matrices on Pose-RAF-DB dataset, **c**-**d** denotes the confusion matrices on Pose-AffectNet dataset



(a) pose > 30°

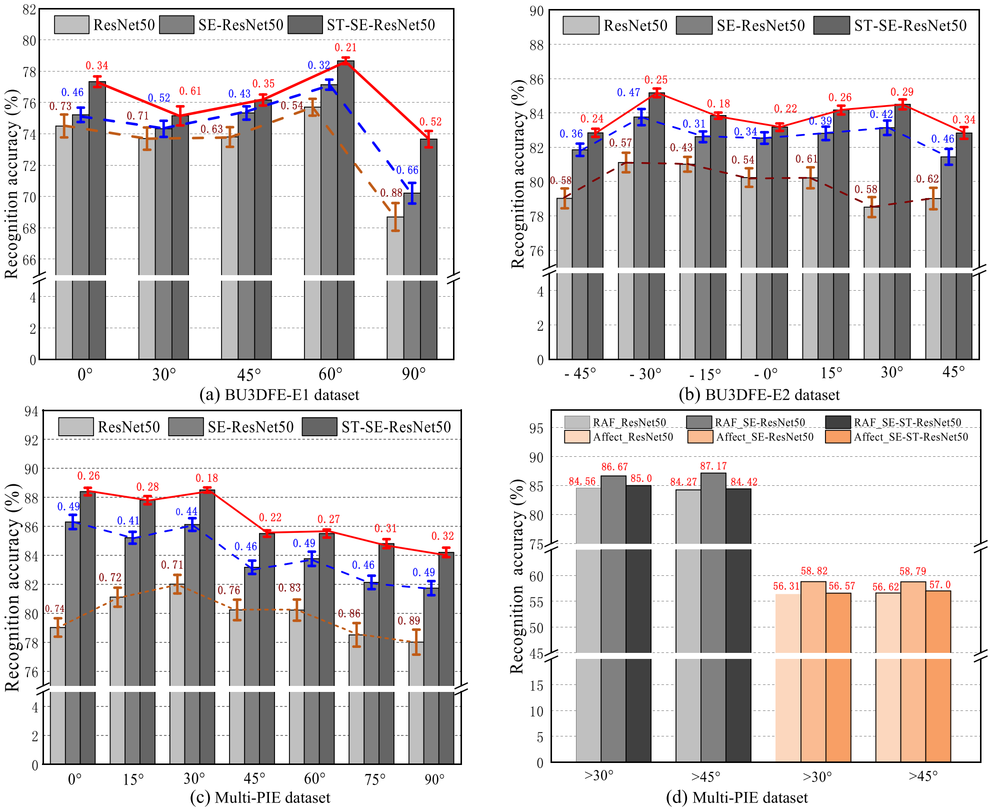
(b) pose > 45°

(c) pose > 30°

(d) pose > 45°

current view and suppress the influence of intensity, pose, background and so on, which are necessary for pose-invariant FER.

As for the database of different scenarios, we also compared the performance of ResNet50, SE-ResNet50 and ST-SE-ResNet50 in controlled and real-world scenarios. The detailed results are provided in Fig. 8a–d, and the corresponding average recognition accuracies can be referred toTables 4, 5, 6, 7, and 8. For the controlled setting, it can be observed that under the influence of different views, the performance of the three models is globally consistent, among which ST-SE-ResNet50 is the lowest, followed by SE-ResNet50, and ResNet50 is the lowest. The stand deviation (SD) of fivefold cross-validations indicates that ST-SE-ResNet50 provides more stable identification accuracy than ResNet50 and SE-ResNet50. This phenomenon is even more pronounced in



(a) BU3DFE-E1 dataset

(b) BU3DFE-E2 dataset

(c) Multi-PIE dataset

(d) Multi-PIE dataset

Fig. 8 a-d the accuracies of Resnet50, SE-Resnet50 and ST-SE-Resnet50 on BU3DFE, Multi-PIE dataset, Pose-FAF-DB and Pose-AffectNet datasets

the Multi-PIE dataset, where the minimum SD values of yaw angle is above 0.7, while that of SE-ResNet50 and ST-SE-ResNet50 is only 0.49 and 0.18, respectively, illustrating that ST-SE block can enhance the stability of the network structure, and it is more robust for pose-invariant FER. For the real-world settings, their recognition accuracy also maintains the same trend, among which ResNet50 is the lowest, followed by ST-SE-ResNet50, and then SE-ResNet50. ST-SE-ResNet50 performs slightly better than ResNet50 on Pose-FAF-DB (0.44% (> 30°), 0.15% (> 45°)) and Pose-AffectNet (0.26% (> 30°), 0.08% (> 45°)) database while compared with the SE-ResNet50, which reduced by (1.67% (> 30°), 2.75% (> 45°)) and (2.25% (> 30°), 1.79% (> 45°)). This result shows that the algorithm in this paper cannot achieve good results in databases with non-normalized poses.

For the influence of expressions intensities, as shown in Tables 2 and 4, the recognition accuracy on the BU3DFE-E1 dataset is much lower than that of the second ones, which can be attributed to the micro-deformation of the low intensities expressions and even more variable yaw angles. In order to illustrate the impact of intensity on facial expression recognition, the classification accuracy of the ST-SE-Resnet50 on the BU3DFE-E1 dataset is shown in Fig. 9. As described in Sect. 4.1, the BU3DFE-E1 dataset contains four different expression intensities. It can be seen from Fig. 9 that the accuracy of expression recognition improves with the intensity level. For the III and IV intensity levels, these textures of the six basic expressions are more obvious than those of low intensity. In this case, the high-level intensity expression images contain more powerful representation capabilities than I and II level intensity. Therefore, the recognition rate of these three methods on BU3DFE-E1 dataset is higher than that on BU3DFE-E2 dataset.

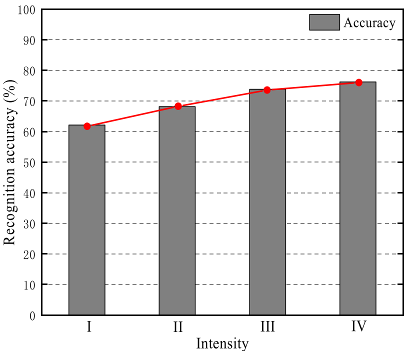


Fig. 9 Influence of four intensities on BU3DFE-E1 dataset

For the misclassified emoticons in the experiment, it is closely correlated to the facial expression image texture. As described in [39], each type of expression can be expressed as a combination of a similar type of textures. When two expressions contain the same type of textures they are more likely to be misclassified. As shown in Figs. 4f and 5h, anger and sadness have a high probability of being misclassified in BU-3DFE dataset, while anger and squint have a high probability of being misclassified in the pose-invariant dataset. This may due to the fact that these expressions include more similar types of textures in their datasets, which can be found in Fig. 3a and c. When the texture types of expressions are notable, the probability of misclassification is relatively low.

For different views, the best recognition results remain between −60° and 60°. In the experiment, when the views are larger than 60°, the recognition accuracy decreases sharply, especially in ResNet50. The reason is that as the view rotate, the main regions of interest (such as eyes, mouth and chin) are gradually blocked, which will reduce the accuracy of recognition. In addition, as can be seen from Tables 3, 5 and 7, the optimal recognition angle is usually not 0°, and they tend to concentrate on near-frontal views. For frontal face images, most of them are symmetrical images, that is to say, half or more than half images can represent the characteristics of the entire expression image. On the contrary, the entire expression images often include much redundant features compared with near-frontal expression images. Therefore, a small yaw angle can not only preserve the frontal facial features, but also add some detailed feature information on the side, which may be more conducive to the task of expression classification.

5 Conclusions

In this paper, a ST-SE-ResNet50 network base on ST-SE blocks was proposed for pose-invariant FER. Herein, the GAP was employed to flatten the feature map into a 1D vector, and then the flattened feature maps were sent to SE block to filter out salient information. The absolute value GAP multiplied SE operation can be regarded as a self-attention mechanism, whose purpose is to force the network to pay more attention to the feature information in the current view and reduce the influence of pose and occlusion on the recognition accuracy. The proposed method was evaluated on four famous datasets, i.e., BU-3DFE, Multi-PIE, Pose-RAF-DB and Pose-AffectNet, and the results indicate that the method is superior to many previous methods in controlled scenarios. However, in the real-world scenario, especially the facial images with different horizontal and pitch angles, the change of recognition accuracy are not obvious relative to the backbone architecture.

Acknowledgements National Natural Science Foundation of China (No:31872399), Advantage Discipline Construction Project (PAPD, No.6-2018) of Jiangsu University

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Shu, X., Yang, J., Yan, R.: Expansion-squeeze-excitation fusion network for elderly activity recognition. arXiv e-prints (2021).
2. Gogić, I., Manhart, M., Pandžić, I.S.: Fast facial expression recognition using local binary features and shallow neural networks. Vis. Comput. **36**(1), 97–112 (2020)
3. Shu, X., Tang, J., Li, Z.: Personalized age progression with bi-level aging dictionary learning. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 905–917 (2018)
4. Shu, X., Tang, J., Li, Z.: Personalized age progression with aging dictionary. In: IEEE International Conference on Computer Vision (ICCV), pp. 3970–3978 (2015).
5. Kumar, S., Bhuyan, M.K., Iwahori, Y.: Multi-level uncorrelated discriminative shared Gaussian process for multi-view facial expression recognition. Vis. Comput. **37**(1), 143–159 (2021)
6. Goh, K.M., Ng, C.H., Li, L.L.: Micro-expression recognition: an updated review of current trends, challenges and solutions. Vis. Comput. **36**(3), 445–468 (2020)
7. Zhu, X., Chen, Z.: Dual-modality spatiotemporal feature learning for spontaneous facial expression recognition in e-learning using hybrid deep neural network. Vis. Comput. **36**(4), 743–755 (2019)
8. Hu, M., Ge, P., Wang, X.: A spatio-temporal integrated model based on local and global features for video expression recognition. Vis. Comput. 1–18 (2021)
9. Zhang, W., Zhang, Y., Ma, L., Guan, J., Gong, S.: Multimodal learning for facial expression recognition. Pattern Recogn. **48**(10), 3191–3202 (2015)
10. Zheng, W.: Multi-view facial expression recognition based on group sparse reduced-rank regression. IEEE Trans. Affect. Comput. **5**, 71–85 (2014)
11. Moore, S., Bowden, R.: Local binary patterns for multi-view facial expression recognition. Comput. Vision. Image Underst. **115**(4), 541–558 (2011)
12. Zhang, F.F., Mao, Q.R., Shen, X.J.: Spatially coherent feature learning for pose-invariant facial expression recognition. ACM Trans. Multimed. Comput. Commun. **14**(1), 1–19 (2018)
13. Liu, Y., Duanmu, M.X., Huo, Z.: Exploring multi-scale deformable context and channel-wise attention for salient object detection. Neurocomputing **428**, 92–103 (2021)
14. Liu, Y., Wei, D., Fang, F., et al. : Dynamic multi-channel metric network for joint pose-aware and identity-invariant facial expression recognition. Inf. Sci. **578**, 195–213 (2021)
15. Liu, Y., Zeng, J., Shan, S.: Multi-channel pose-aware convolution neural networks for multi-view facial expression recognition. In: 13th IEEE International Conference on Automatic Face & Gesture Recognition, pp. 458–465 (2018).
16. Zhang, K., Huang, Y., Du, Y.: Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Trans. Image Process. **26**, 4193–4203 (2017)
17. Zheng, H., Wang, R., Ji, W.: Discriminative deep multi-task learning for facial expression recognition. Inf. Sci. **533**, 60–71 (2020)
18. Ma, H., Celik, T., Li, H.C.: Lightweight attention convolutional neural network through network slimming for robust facial expression recognition. SIViP. **15**(7), 1507–1515 (2021)

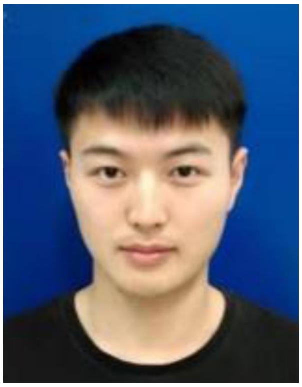
19. Li, Y., Lu, G., Li, J.: Facial Expression Recognition in the Wild Using Multi-level Features and Attention Mechanisms. IEEE Trans. Affect. Comput. **10**(99), 1–1 (2020)
20. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell. **42**(8), 7132–7141 (2017)
21. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of 32nd International Conference on Mechanical Learning, Lille, France, pp. 448-456 (2015).
22. Lin, M., Chen, Q., Yan, S.: Network in network. In: Proceedings of International Conference on Learning Computer Science. Vol 20, Issue 13, (2014).
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference Computer Vision Pattern Recognit (CVPR). pp. 770–778 (2016).
24. Yin, L., Wei, X., Sun, Y., et al. : A 3D facial expression database for facial behavior research. In: Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition, pp. 211–216 (2006).
25. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-PIE. Image Vis. Comput. **28**(5), 807–813 (2010)
26. Li, S., Deng, W.: Reliable crowdsourcing and deep locality preserving learning for unconstrained facial expression recognition. IEEE Trans. Image Process. **28**, 356–370 (2019)
27. Mollahosseini, A., Hasani, B., Mahoor, M.H.: AffectNet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affect. Comput. **10**, 18–31 (2019)
28. Wu, J.L., Lin, Z.C., Zheng, W.M.: Locality-constrained linear coding based bi-layer model for multi-view facial expression recognition. Neurocomputing **239**, 143–152 (2017)
29. Jung, H., Lee, S., Yim, J.: Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of International Conference Computer Vision pp. 2982–2991 (2015).
30. Zhang, T., Zheng, W., Cui, Z.: A deep neural network-driven feature learning method for multi-view facial expression recognition. IEEE Trans. Multimed. **18**(12), 2528–2536 (2016)
31. Zhang, F., Zhang, T., Mao, Q., Xu, C.: Geometry guided pose-invariant facial expression recognition. IEEE Trans. Image Process. **29**, 4445–4460 (2020)
32. Can, W., Wang, S., Liang, G.: Identity and pose-robust facial expression recognition through adversarial feature learning. In: The 27th ACM International Conference ACM. pp. 238–246 (2019).
33. Jampour, M., Mauthner, T., Bischof, H.: Multi-view facial expressions recognition using local linear regression of sparse codes. In: Computer Vision Winter Workshop Paul Wohlhart (2015).
34. Fan, J., Wang, S., Yang, P., et al. : Multi-view facial expression recognition based on multitask learning and generative adversarial network. In: IEEE International Conference on Industrial Informatics. (2020).
35. Wang, K., Peng, X., Yang, J.: Region attention networks for pose and occlusion robust facial expression recognition. IEEE Trans. Image Process. **29**, 4057–4069 (2020)
36. Gera, D., Balasubramanian, S.: CERN: Compact facial expression recognition net. Pattern Recogn. Lett. **155**, 9–18 (2022)
37. Gera, D., Balasubramanian, S.: Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition. Pattern Recogn. Lett. **145**, 58–66 (2021)
38. Zhao, Z., Liu, Q., Wang, S.: Learning deep global multi-scale and local attention features for facial expression recognition in the wild. IEEE Trans. Image Process. **30**, 6544–6556 (2021)
39. Wang, Z.N., Zeng, F.W.: OAENet: Oriented attention ensemble for accurate facial expression recognition. Pattern Recogn. **112**(5), 107694 (2021)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chaoji Liu received M.S. degree from the School of Mechanical and Electrical Engineering, Tarim University, China in 2018. He is currently a PhD student in Jiangsu University, School of Electrical Information Engineering, China. His current research interests focus on image processing, computer vision and deep learning.

Xingqiao Liu received the Ph. D. degree from Jiangsu University, School of Electrical Information Engineering, China in 2009. From May to December 2007, he was a visiting scholar in the Laboratory of Power Electronics and Electric Drives, SHEFFIELD University, UK. His research interests include machine intelligence, pattern recognition, and their applications on human motion analysis, multi-fingered robotic hand control, human-robot interaction and collaboration, and robot skill learning. And current research interests include image processing, hardware-based image processing and multimedia data processing and analysis.

Chong Chen Associate Professor, he received PhD in Jiangsu University, School of Electrical Information Engineering, China in 2021. His current research interests focus on image processing, hardware-based image processing and multimedia data processing and analysis.



Qiankun Wang is currently pursuing M.S. degree in Jiangsu University, School of Electrical Information Engineering, China in 2021. His research interests include image processing, video surveillance and multimedia data processing.