International Journal of Machine Learning and Cybernetics (2024) 15:5863–5880
https://doi.org/10.1007/s13042-024-02287-0

ORIGINAL ARTICLE

A dual stream attention network for facial expression recognition in the wild

Hui Tang[1] · **Yichang Li**[2] · **Zhong Jin**[1,2]

Received: 3 September 2023 / Accepted: 14 July 2024 / Published online: 23 July 2024

Abstract
Facial Expression Recognition (FER) is crucial for human-computer interaction and has achieved satisfactory results on lab-collected datasets. However, occlusion and head pose variation in the real world make FER extremely challenging due to facial information deficiency. This paper proposes a novel Dual Stream Attention Network (DSAN) for occlusion and head pose robust FER. Specifically, DSAN consists of a Global Feature Element-based Attention Network (GFE-AN) and a Multi-Feature Fusion-based Attention Network (MFF-AN). A sparse attention block and a feature recalibration loss designed in GFE-AN selectively emphasize feature elements meaningful for facial expression and suppress those unrelated to facial expression. And a lightweight local feature attention block is customized in MFF-AN to extract rich semantic information from different representation sub-spaces. In addition, DSAN takes into account computation overhead minimization when designing model architecture. Extensive experiments on public benchmarks demonstrate that the proposed DSAN outperforms the state-of-the-art methods with 89.70% on RAF-DB, 89.93% on FERPlus, 65.77% on AffectNet-7, 62.13% on AffectNet-8. Moreover, the parameter size of DSAN is only 11.33M, which is lightweight compared to most of the recent in-the-wild FER algorithms.

Keywords Facial expression recognition · Occlusion and head pose variation · Sparse attention block · Local feature attention block · Lightweight model

1 Introduction

Facial Expression (FE), one of the most potent, natural, and immediate means for human beings to carry their emotional states and intentions [1–3], has been extensively studied in various active research fields. Recently, with the widespread application of human-computer interaction (e.g., social robots, driver fatigue monitoring, intelligent tutoring systems, and mental health analysis), automatic facial expression analysis received increasing attention in the fields of computer vision and machine learning [4].

In the past decades, automatic Facial Expression Recognition (FER) systems [5, 6] have performed perfectly in lab collected databases, such as CK+ [7], Oulu-CASIA [8], and MMI [9]. However, due to the uncertainties in the wild, the performance of FER methods trained in the constrained environment degrades dramatically. As shown in Fig. 1, facial occlusion, head pose variation, and the compound of both are prevalent in wild FER datasets, causing significant changes to facial appearance. Furthermore, facial appearances of various expressions are inherently similar, and non-facial regions would aggravate the intra-class differences and inter-class similarities, making facial expressions challenging to classify [10]. Hence, facial occlusion and head pose variation for FER in the wild are urgent problems to be solved.

Recently, several large-scale In-The-Wild (ITW) datasets (such as RAF-DB [11], FERPlus [12], and AffectNet [13]) along with some benchmarks have been proposed to

✉ Zhong Jin
zhongjin@njust.edu.cn

Hui Tang
220106011114@njust.edu.cn

Yichang Li
liyic@cupk.edu.cn

[1] Department of Computer Science and Engineering, Nanjing University of Science and Technology, No. 200 Xiao Lingwei Street, Nanjing 210094, Jiangsu, China

[2] Department of Computer, China University of Petroleum-Beijing at Karamay, No. 255 Anding Road, Karamay 834000, Xinjiang, China

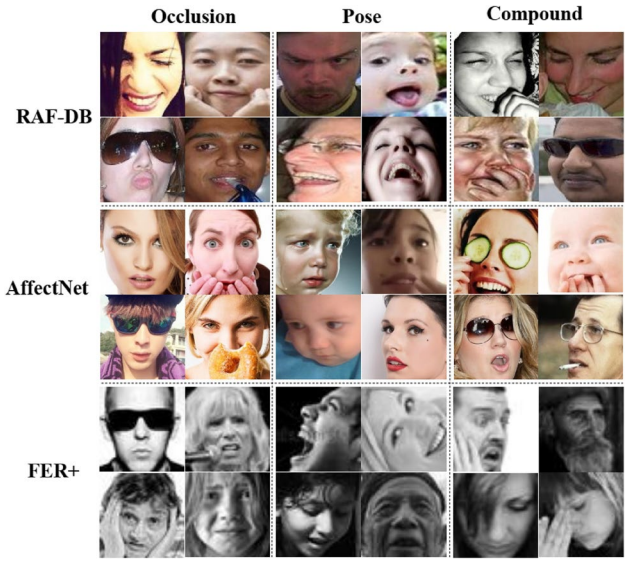


Fig. 1 The samples from RAF-DB, AffectNet, and FER+ datasets. Face occlusions, variant head poses, and the compound of both can be seen in the above images, which bring trouble to FER

enhance the progress of FER in real-world scenarios. As a widely studied technology in computer vision, the attention mechanism is used to solve uncertainties in the wild. Some approaches separate the face into several regions and then weight each region by an attention mechanism to alleviate the effect of occlusions and variant poses. However, the regions are usually derived according to facial landmark points [14, 15], or cropping [15, 16], which may result in misalignment or uncertainty for FER. Zhao et al. [17] tried to alleviate the uncertainty in the wild by dividing the pre-extracted feature map into non-overlapping local regions, but it introduced additional spatial attention parameters, which affected the portability of the algorithm. Deeper convolution has been proven to produce more discriminative features for computer vision [18], but it is sensitive to occlusion and pose variation [17], which will enlarge the uncertainty of FER. To make use of the powerful representation ability of deeper convolutions and not be affected by occlusion and pose variation, several researchers [10, 19] attempt to learn robust feature representation by multi-scale networks. Ma et al. [20] introduced a model called visual transformers with feature fusion (VTFF), which designed an attentional feature fusion model to integrate LBP and CNN features adaptively to enrich the representation of the visual words and the attention. Sun et al. [21] also explored texture enhancement block by extracting the LBP and gray-level co-occurrence matrix features, providing more textural information about facial images. In addition, Wang et al. [22] and Li et al. [23] tacked ambiguous labels and class imbalance to improve the recognition capability of FER in an uncertain environment. However, the above methods achieved effective performance with excessive parameters and FLOPs, which lead to deployment obstacles in practical applications (e.g., smartphones, sensors, and handheld devices). In recent years, lightweight networks, such as Double Dynamic Relationships Graph Convolutional Network (DDRGCN) [24] and FaceNet2ExpNet [25], have made satisfactory results in lab-collected databases. However, it is difficult for lightweight models to learn efficient facial expression features due to the uncertainties of the large-scale FER databases collected from realistic scenarios [26–28]. Therefore, enhancing the performance of FER in the wild while saving computing overhead is a meaningful and unsolved matter.

To this end, this paper proposes a novel Dual Stream Attention Network (DSAN) for pose and occlusion robust FER. The DSAN mainly comprises two crucial self-attention networks: a Global Feature Element-based Attention Net (GFE-AN) and a Multi-Feature Fusion-based Attention Net (MFF-AN). Unlike previous works that apply attention to landmark point regions or cropping blocks, our GFE-AN allows the network to perform feature recalibration without landmark guidance. It performs an attention mechanism on the final feature vector to selectively emphasize informative feature elements and suppress those irrelevant to facial expression, improving the robustness of the model to occlusion and pose variation. Furthermore, to further exploit the representation ability of deep convolutions and to keep it insensitive to occlusion and pose variation, inspired by multi-head attention [29], we propose a lightweight yet efficient attention module MFF-AN to extract various facial features from multiple networks. The sparse attention in GFE-AN can efficiently alleviate occlusion and pose variation problems but suppress partial facial expression information. MFF-AN extracts rich semantic information for the original facial image, which makes up for the deficiency of GFE-AN.

As shown in Fig. 2, we first use a convolutional neural network to extract basic features. Then, we build a GFE-AN network to recalibrate the feature elements, where a feature recalibration loss is applied to increase the inter-class distance and decrease the intra-class distance. Meanwhile, we design MFF-AN to concurrently capture complementary local facial features, where each local feature in sub-networks tends to learn richer semantic information. The output of multiple sub-networks in MFF-AN will be weighted and summed into a comparatively comprehensive feature representation. Finally, the two branches are trained in parallel to save computing time, and the class confidences of the two branches will be combined to predict the expression label.

In summary, the main contributions of the proposed work are as follows:

- A novel Dual Stream Attention Network (DSAN) is proposed for facial expression recognition in the wild. DSAN can adaptively acquire the importance

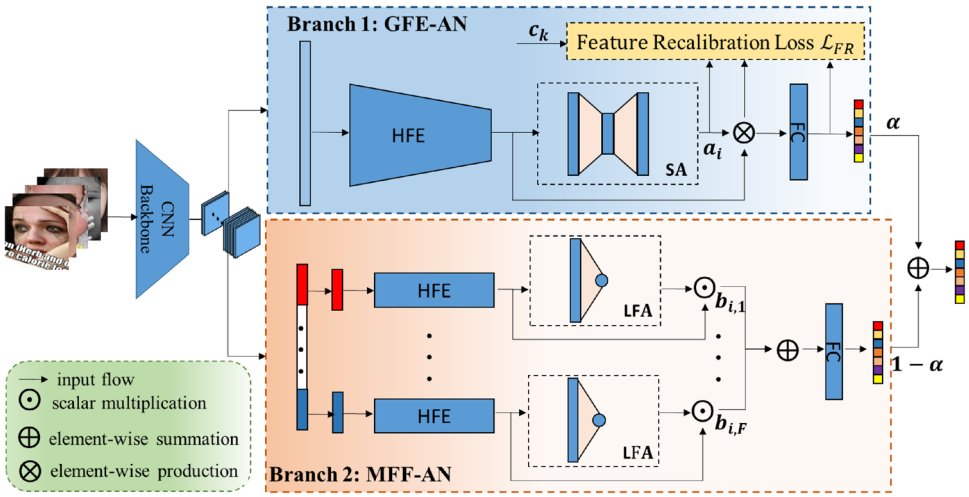


Fig. 2 An overview of our proposed DSAN. The method consists of two components, including a Global Feature Element-based Attention Net (GFE-AN) and a Multi-Feature Fusion-based Attention Net (MFF-AN). The former branch obtains the importance coefficient (weight) of each element of the global feature in calculating feature recalibrate loss. The later branch fusion the feature representation learned from different sub-networks to alleviate the sensitivity of the deep network to facial occlusion and pose variation. HFE is a high-level feature extractor. SA and LFA represent sparse attention and local feature attention, respectively. The prediction expression label from both branches jointly determines the final classification result

coefficients of global feature elements and local features with few parameters and FLOPs, which is robust for both occlusion and pose variation.

- We design a simple yet efficient GFE-AN module to rearrange the global feature and then execute intra-class aggregation and inter-class separation. Notably, under the supervision of feature recalibration loss, the expression-related features will be emphasized, and those in the opposite direction will be suppressed.
- We propose an MFF-AN module to aggregate multiple local facial features with complementary information. It enhances the recognition capability of facial feature expression and alleviates the susceptibility of the deeper network toward local non-facial regions.
- The experiment results of our DSAN on three publicly available FER in the wild datasets show outstanding performance compared to the state-of-the-art methods. Moreover, experiments on realistic occlusion and pose variation test datasets are also performed excellently.

The rest of this article is organized as follows. Section 2 briefly reviews related literature of FER in two respects. The details of our proposed DSAN are presented in Sect. 3. Section 4 presents extensive experimental evaluation results and analysis. The discussion and the conclusion are given in Sects. 5 and 6.

2 Related work

With the development of deep learning and the release of large-scale ITW FER databases, extensive efforts have been made for FER. This section reviews the previous works related to our DSAN from two aspects: FER in the wild and attention mechanism.

2.1 FER in the wild

Generally, FER can be divided into handcrafted features-based methods and deep learning-based methods. Local Binary Pattern (LBP) [30–32], Gabor wavelet [33, 34], Local Phase Quantization (LPQ) [35], Scale-Invariant Feature Transform (SIFT) [36], and Histogram of Oriented Gradients (HOG) [37] are the mainstream feature extractors of handcrafted features-based methods. Gradually, due to the rapid progress of GPU units and the completion of large-scale datasets, various deep learning methods were proposed for FER and exceeded previous works by a larger margin. Umer et al. [38] applied some novel data augmentation techniques to improve the representation ability of a network and use the network to learn rich facial features. Zhao et al. [39] improved the

model's generalization ability by reducing the difference between peak and non-peak expressions, which made it easier for the model to recognize hard samples. Auto FER on lab-collected datasets has achieved great performance in the past few decades. However, it suffered a bottleneck in the realistic scenario because of the uncertainties, including blurred pictures, label noise, occlusion, and pose variation. FER in the wild has received extensive attention with the development of ITW datasets. We will briefly introduce several FER algorithms according to the mentioned uncertainties.

2.1.1 Super-resolution

Super-Resolution (SR) is an effective solution for blurred pictures, which recovers the corresponding higher-resolution images from the observed low-resolution images [40]. Shao et al. [41] designed a three-stream SR network for FER, which reconstructed SR image by weighting and summing the outputs of three branches: edge enhancement network, upsampling network, and SR primary network, making tiny facial expressions more distinguishable. However, although the recognition accuracy was performed well, the images after SR reconstruction still appear blurry to the human eyes. Aiming at this problem, Than et al. [42] developed a pyramid with a super-resolution network for FER in the wild, and the EDSR [43] architecture which was used to upscale the image size for the super-resolution task.

2.1.2 Noise label

For noise labels, the Inconsistent Pseudo Annotations to Latent Truth (IPA2LT) [44] framework was the first method to train a FER model from multiple inconsistently labeled datasets and large-scale unlabeled data. It tagged multiple labels for each image and then learned a FER model to fit the underlying truth. Subsequently, a self-cure network [22] was built to suppress the uncertainties caused by crowdsourcing and ambiguous facial images in the wild. Moreover, Li et al. [23] introduced an adaptive importance coefficient module to reweight the contribution of the tail expression representation, which alleviated the influence of class imbalance.

2.1.3 Occlusion and pose variation

Occlusion and pose variation are the main obstacles of FER in the nature scene. As shown in Fig. 1, the facial region is easily obscured by hair, sunglasses, and hands, etc. So far, networks based on attention are the better option to figure out those problems, and we will introduce a few approaches in the following subsection. In addition, the ITW dataset contains a wide variety of facial images because the sample acquisition environment is not controlled in the wild. Using only the cross-entropy loss function as supervision does not yield satisfactory results. Wen et al. [45] proposed a center loss function to push intra-class compactness as much as possible. Cai et al. [46] extended the center loss to island loss by reducing the pairwise cosine distance between class centers in feature space. A Deep Attentive Center Loss (DACL) [47] reformulated the center loss as sparse center loss and combined it with an attention mechanism. It adaptively selects a subset of significant feature elements for enhanced discrimination. Li et al. [48] designed a separate loss based on normalized cosine to enhance the discriminability of the learned features. Note the last three loss functions are dedicated to intra-class aggregation and inter-class separation, and the sparse center loss performs best due to the selectivity of the attention mechanism. In contrast, our recalibration loss utilizes feature recalibration to reformulate the sparse center loss function. It can obtain more discriminative sparse features when implementing intra-class aggregating and inter-class separation.

2.2 Attention mechanism

The attention mechanism is a model that simulates the attentional mechanism of the human brain [49]. It can be considered a combinatorial function that highlights the effect of the critical input on the output by calculating the probability distribution of attention. Recently, the attention mechanism has been widely applied to FER in the wild. To fully use the relationship between global features, Chen et al. [50] integrates channel attention to 3D-Inception-ResNet [51] to learn the saliency mapping of spatiotemporal features along different dimensions. Wu et al. [52] introduced local and global attention in cascade expression focal GAN, where global attention captures the most salient expression changes (i.e. the mouth region) and local attention captures imperceptible expression changes at local region. Similarly, Li et al. [14] proposed a Convolution Neural Network with an Attention (ACNN) mechanism for FER to tackle the face occlusion challenge. Patch-based ACNN (pACNN) and global–local-based ACNN (gACNN) are the two versions of ACNN. The former focuses on local discriminative and representative patches, and the latter helps infer local details and global context cues. Region Attention Network (RAN) [15] is also an attention network designed for occlusion robust. Compared to the ACNN, RAM refines the attention weights with a relation-attention module and region bias loss function. However, the above algorithms divide a global image or feature maps into multiple local patches, resulting in misalignment or uncertainty to FER in the wild. To solve this problem, Zhao et al. [17] employed a local attention module to divide the mid-level feature maps into several local feature maps without overlap. Each local feature map can autonomously

focus on salient local features by an attention mechanism. Subsequently, A novel cross-modality attention fusion network [53] was developed to fuse the features from different facial modalities, which could enhance the spatial correlations between different facial modalities. Unlike the above attention networks, reference [54] proposed a novel Spatial-Channel Attention Net (SCAN) that obtained both local and global attention per channel per spatial location. SCAN made the FER model more robust to occlusion and pose variation without seeking any information from the landmark detector. Our DSAN is also a model independent of landmark guidance, which needs less computing overhead than other attention-based methods.

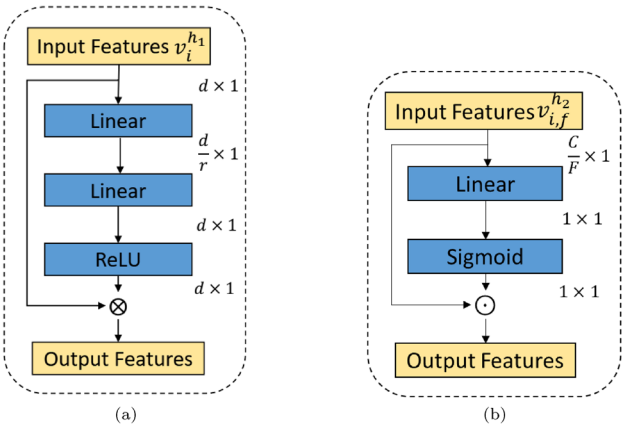


Fig. 3 Two types of attention blocks are employed in our DSAN. **a** Sparse attention (SA) block. **b** Local feature attention (LFA) block

3 Method

In this section, we first show an overview of the Dual Stream Attention Network (DSAN) for facial expression recognition in the wild. Then, we explicitly illustrate its two main branches: the Global Feature Element-based Attention Network (GFE-AN) and the Multi-Feature Fusion-based Attention Network (MFF-AN). Finally, we give a detailed description of the fusion strategy and loss function.

3.1 Overview

As shown in Fig. 2, DSAN is constructed upon traditional CNNs and mainly consists of two parallel modules, including GFE-AN and MFF-AN. Given a batch of facial expression samples, we first fed them into the CNN backbone (ResNet-18 [55]) to extract feature maps and then convert them into the basic feature vector by using max-pooling. GFE-AN extracts the higher-level feature, and the sparse attention block adaptively assigns a weight for each element of the higher-level feature. We then design a feature recalibration loss to enhance the importance of the elements related to facial expression and suppress those unrelated to facial expression. Subsequently, we decompose the basic feature vector into multiple local features and extract their high-level semantic information in the MFF-AN module. Each high-level feature learns a coefficient from itself and then weights it into the final feature representation. Finally, the class confidences from both branches jointly determine the final recognition result. The details of each module are described as follows.

3.2 Global feature element-based attention network

As discussed earlier, face images in the natural scene tend to contain non-facial areas, resulting in a large number of redundant elements in the feature vector. Therefore, we build a global feature element-based attention network to selectively emphasize feature elements meaningful for expression classification and suppress those irrelevances. Specifically, given a batch of N training samples $\{(x_i, y_i)|i = 1, \ldots, N\}$, where x_i is the i-th input sample and $y_i \in \{1, \ldots, K\}$ is its corresponding label for the K-class FER task. The sample is fed into the CNN backbone to obtain feature maps F_i with the size of $C \times W \times H$, where C, W, and H represent the number of channels, height, and width of feature maps, respectively. Max-pooling retains the original value in a region in the backward phase, and the backward result of max-pooling is a sparse matrix that favors the model to converge [56]. So we use max-pooling to extract the max value along the spatial dimension of F_i, the result of which is a feature vector $v_i \in \mathbb{R}^{C\times 1}$. Notably, CNN learns increasingly discriminative features as the model's hierarchies become deeper [18], and the number of parameters of the model will increase dramatically if the dimension of v_i is too large. To balance the accuracy and computing overhead, we employ a linear Fully-Connected (FC) layer and a ReLU activation function σ_1 to build a high-level feature extractor. It maps v_i to a low dimensional space. This operation can be formulated as:

$$v_i^{h_1} = \sigma_1\left(\mathbf{W}^{h_1\mathrm{T}} v_i + \mathbf{b}^{h_1}\right) \tag{1}$$

where $\mathbf{W}^{h_1} = [w_1^{h_1}, w_2^{h_1}, \ldots, w_d^{h_1}] \in \mathbb{R}^{C\times d}, d \le C$ and $\mathbf{b}^{h_1} = [b_1^{h_1}, b_2^{h_1}, \ldots, b_d^{h_1}] \in \mathbb{R}^{d\times 1}$ denote the weights and bias of the FC layer used for extracting the d-dimension high-level feature vector for i-th sample in the first branch, respectively.

To address facial occlusion and pose variation in FER in the wild, we work on weighting each feature element to recalibrate v_i, reducing the intra-class distance while increasing the inter-class variance. Inspired by the lightweight block Squeeze-and-Excitation (SE) [57] and DACL [47], we customize a Sparse Attention (SA) with

bottleneck structure to assign weights through global information. As shown in Fig. 3a, the attention block contains two stacked FC layers, and a ReLU activate function where r is the reduction ratio to control the size of the model parameters. To investigate the trade-off between performance and computing overhead, we explore the impact of r in Sect. 4.4.6. The sparse attention wights for $v_i^{h_1}$ are calculated as:

$$a_i = \sigma_1\left(\mathbf{W}_2^{sa\mathrm{T}}(\mathbf{W}_1^{sa\mathrm{T}} v_i^{h_1} + \mathbf{b}_1^{sa}) + \mathbf{b}_2^{sa}\right), \tag{2}$$

where $\mathbf{W}_1^{sa} \in \mathbb{R}^{d\times\frac{d}{r}}$, $\mathbf{b}_1^{sa} \in \mathbb{R}^{\frac{d}{r}\times 1}$ and $\mathbf{W}_2^{sa} \in \mathbb{R}^{\frac{d}{r}\times d}$, $\mathbf{b}_2^{sa} \in \mathbb{R}^{d\times 1}$ are the weights and biases for the first and second FC layer in the sparse attention network. σ_1 is the ReLU activation function that conduces to the sparsity of a_i and converges much faster than Sigmoid and Tanh.

Using cross-entropy loss as the objective function for FER in the wild is insufficient to produce discriminative features. Here, we develop a feature recalibration-based cross-entropy loss $\mathcal{L}_{FR\_CE}$ to weigh the feature elements in $v_i^{h_1}$ so that the feature vector is recalibrated.

$$\begin{aligned}\mathcal{L}_{FR\_CE} &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_i \log P\left(y_i = k|a_i \otimes v_i^{h_1}\right) \\ &= -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\mathbf{W}_{y_i}^{p\mathrm{T}}(a_i\otimes v_i^{h_1})+\mathbf{b}_{y_i}^{p}}}{\sum_{k=1}^{K} e^{\mathbf{W}_k^{p\mathrm{T}}(a_i\otimes v_i^{h_1})+\mathbf{b}_k^{p}}},\end{aligned} \tag{3}$$

where $\mathbf{W}^p$ and $\mathbf{b}^p$ are the weights and bias for the FC layer, both of which are used to obtain the label probability of y_i on the prediction stage, $\otimes$ means element-wise multiplication. Under the supervision of $\mathcal{L}_{FR\_CE}$, the SA block can selectively assign weights for elements in the feature vector, emphasizing informative feature elements related to facial expression and suppressing those unrelated ones.

As described in Sect. 2.1, the sparse center loss has achieved a favorable result, but the class center of which is not precise due to the effects of occlusion and pose variation. To solve this problem, we reformulate the sparse center loss [47] as a novel feature recalibration-based center loss function:

$$\mathcal{L}_{FR\_C} = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{d} a'_{i,j}|a_{i,j}v_{i,j}^{h_1} - c_{y_i,j}|^2, \tag{4}$$

the subset of attention weight a_i is obtained by optimizing $\mathcal{L}_{FR\_CE}$ and used to recalibrate the feature elements of $v_i^{h_1}$, which can effectively alleviate the negative influence of non-facial regions for FER in the wild. a'_i plays the same role as the attention weights in the sparse center loss. It aims to select only a subset of elements that make all samples of the same class closest to the class center. a'_i and a_i share the same value. Since the GPU capacity is limited, it is not feasible to compute the class center features for all training samples. Following reference [45–47], we initially define the class-center features and then dynamically extract the class-center features during the training process by minimizing equation (). The feature recalibration-based center loss selectively penalizes the distance between the feature element and its corresponding class center at each dimension. The model learns a relatively accurate class center for samples of the same category without the influence of non-facial regions.

$\mathcal{L}_{FR\_CE}$ and $\mathcal{L}_{FR\_C}$ jointly supervise DSAN tend to reduce the intra-class variation of the feature vector. However, the data collected in the wild contain some confusing samples, which result in the overlap between the clusters of different classes. Here, we introduce an island loss [46] term to expand the pairwise distance between class centers in the feature space. We rename it as class center separation loss, which is defined as:

$$\mathcal{L}_{CCS} = \sum_{c_i\in K}\sum_{\substack{c_j\in K\\ c_j\neq c_i}}\left(\frac{c_i\cdot c_j}{\|c_i\|_2\|c_j\|_2}+1\right), \tag{5}$$

where $(\cdot)$ represents the dot production. c_i and c_j denote the i-th and j-th class center, respectively.

It can be seen that the above three loss functions are proposed based on feature recalibration. Hence, we call the overall loss function of the first stream GFE-AN network as feature recalibration loss $\mathcal{L}_{FR}$. The mathematical formula of $\mathcal{L}_{FR}$ is defined as:

$$\mathcal{L}_{FR} = \mathcal{L}_{FR\_CE} + \lambda_1\mathcal{L}_{FR\_C} + \lambda_2\mathcal{L}_{CCS}, \tag{6}$$

where λ_1 and λ_2 are used for balancing the three terms of $\mathcal{L}_{FR}$. Significantly, Eq. (6) can be reduced to the standard cross-entropy $\mathcal{L}_{CE}$, center loss $\mathcal{L}_C$, and island loss $\mathcal{L}_I$, when $a_{i,1} = a_{i,2} = \cdots = a_{i,d}$. The details can be written as:

$$\mathcal{L}_{FR} = \begin{cases}\mathcal{L}_{CE}, & \text{if } \lambda_1 = \lambda_2 = 0\\ \mathcal{L}_C, & \text{if } \lambda_2 = 0\\ \mathcal{L}_I, & \text{otherwise.}\end{cases} \tag{7}$$

To demonstrate the effectiveness of our feature recalibration loss, we give the visualization feature and ablation study in Sect. 4.4.2.

3.3 Multi-feature fusion-based attention network

Notably, feature elements with greater weight contribute more to expression classification. The weights of some feature elements that interfere with FER are set to 0 to deal with facial occlusion and pose variation. We measure the proportion of

the removed feature elements in a sample with the average value of $a_{i,j} = 0$ (expressed as $P_{a_{i,j}=0}$). We found that when using only GFE-AN, $P_{a_{i,j}=0}$ is greater than 0.5, which means that information related to facial expressions can also be suppressed. To supplement the lost facial information for GFE-AN, we introduce another branch called the multi-feature fusion-based attention network to learn rich facial feature information from multiple sub-networks.

As shown in Fig. 2, the feature vector v_i be split into multiple local features $v_{i,f} \in \mathbb{R}^{\frac{C}{F}\times 1}$ firstly, where $f \in [1, \ldots, F]$ and F is the number of local feature. Then the $v_{i,f}$ is fed into different sub-network to extract high-level local feature $v_{i,f}^{h_2}$ concurrently by Eq. (1), where $\mathbf{W}^{h_2} \in \mathbb{R}^{\frac{C}{F}\times\frac{C}{F}}$ and $\mathbf{b}^{h_2} \in \mathbb{R}^{\frac{C}{F}\times 1}$ in this branch. Subsequently, the high-level local features learn a weight $b_{i,f}$ from themselves by Local Feature Attention (LFA), which is shown in Fig. 3b. Finally, all high-level local features are weighted and summed instead of directly concatenated, which is helpful to parameter saving for the subsequent operations and extracts more discriminating features.

$$b_{i,f} = \sigma_2\left(\mathbf{W}^{lfa\mathrm{T}} v_{i,f}^{h_2} + \mathbf{b}^{lfa}\right), \tag{8}$$

where $\mathbf{W}^{lfa} \in \mathbb{R}^{\frac{C}{F}\times 1}$ and $\mathbf{b}^{lfa} \in \mathbb{R}^{1\times 1}$ are the weights and bias of the LFA. σ_2 is Sigmoid activate function, which limits $0 < b_{i,f} < 1$. The final high-level feature in branch two is:

$$v_i^{h_2} = \sum_{f=1}^{F} b_{i,f} \odot v_{i,f}^{h_2}, \tag{9}$$

where $\odot$ means scalar multiplication. MFF-AN adaptively aggregates local features from wider domains, thus acquiring features with better characterization capabilities. It is able to effectively supplement the missing expression information in GFE-AN, which alleviates the sensitivity of the DSAN to local occlusion and pose variation.

MFF-AN is an efficient and lightweight attention model, the parameter number of which is less than that of an FC layer. To verify this idea, we calculate the parameter size for MFF-AN and an FC layer without considering the bias. The number of parameters in HFE and LFA of MFF-AN is $(\frac{C}{F})^2 \times F$ and $\frac{C}{F} \times F$, while that in an FC layer is C^2. Notable, the number of parameters in MFF-AN is less than that in an FC layer when Eq. (10) holds, where $C > 1$.

$$F > \frac{C}{C-1} > 1. \tag{10}$$

We can draw two conclusions from the above analysis. The MFF-AN becomes lighter as F enlarges; The number of parameters of MFF-AN is less than that of an FC layer when $F > 1$.

3.4 Fusion strategy and loss function

Pixel-level fusion, feature-level fusion, and decision-level fusion are three conventional image fusion methods in computing vision tasks [58, 59]. The first technique combines each pixel in source images for further computer processing tasks [60], which is only applicable to the fusion of different modes. The second technique concatenates supplementary feature vectors to construct a joint feature vector, then uses it to train a classifier for FER [61]. The last technique is the highest level among the above three fusion methods, which integrates the recognition results of each modality. The vast majority of studies on bimodal affect recognition are built on the decision-level fusion technique [17]. In this paper, we design two independent models, GFE-AN and MFF-AN. The elements in the feature vector of the GFE-AN model are weighted to focus more on the facial area related to the expression, and the noise elements generated by occlusion and head rotation are also suppressed, but at the same time, some of the facial information is inevitably lost. The MFF-AN module employs multiple sub-networks to extract features, yielding a richer representation of facial expressions, which compensates for the missing facial information in the GFE-AN's representation. Additionally, due to the noise existing in natural scene samples, MFF-AN may introduce some redundant noise information. The GFE-AN module can suppress noise elements, which mitigates the impact of sample noise on the entire model. Therefore, the features of the two branches are complementary, so the proposed DSAN employs a decision-level fusion strategy. The features of the two branches after the FC layer and Softmax, respectively, are weighted and fused to determine the final expression category.

The loss function in GFE-AN is feature recalibration loss represented as $\mathcal{L}_{FR}$, and the loss function in MFF-AN is cross-entropy loss represented as $\mathcal{L}_{CE}$. The final loss function of DSAN is defined as:

$$\mathcal{L} = \alpha\mathcal{L}_{FR} + (1-\alpha)\mathcal{L}_{CE}. \tag{11}$$

where α is a hyper-parameter that balances the two branches of DSAN.

4 Experiments

In this section, we first briefly describe the used ITW FER datasets and implementation details. We then evaluate our proposed method on the wild datasets compared with various state-of-the-art FER approaches. In addition, we

explore the impact of parameters and further verify the effectiveness of the DSAN components. Finally, we have a discussion about the obtained experimental results.

4.1 Datasets

4.1.1 RAF-DB

RAF-DB [11] is a real-world facial affective dataset. It comprised about 30,000 facial images retrieved from the internet. The dataset consists of the basic emotion subset and the compound emotion subset, the image of which was annotated by 40 trained human coders. In the experiment, we use the basic emotion subset with 12,271 training images and 3068 test images, including seven basic expressions (i.e. surprise, fear, disgust, happiness, sadness, anger, neutral).

4.1.2 FERPlus

FERPlus [12], also called FER+, which is the extension of FER 2013 [62] employed in the ICML 2013 challenges in representation learning. The images of FERPlus are all collected by the Google search engine, which contains 28,709 training images, 3589 validation images, and 3589 test images. The images in FERPlus are gray-scale with a size of 48×48, and each image is assigned ten taggers to one of eight categories. FERPlus added a *Contempt* label compared to RAF-DB. For a fair comparison, we report the accuracy of overall samples under the supervision of majority voting for performance measurement.

4.1.3 AffectNet

The AffectNet [13] is by far the largest FER dataset in the wild, with more than 1,000,000 facial images collected from the internet. However, only about 450,000 images have been annotated manually with 11 expression categories. AffectNet contains two benchmark branches, AffectNet-7 (AffectNet with seven classes) and AffectNet-8 (AffectNet with eight classes; the additional class is *Contempt*). For AttectNet-7, there are 283,901 images as training data and 3500 images as test data. For the AffectNet-8, there are 287,568 images as training data and 4000 images as test data.

4.1.4 Occlusion and pose variation datasets

To verify the robustness of our algorithm to occlusion and pose variation, we test our model on several test subsets with occlusion and pose annotation (e.g., Occlusion-RAF-DB, Pose-RAF-DB, Occlusion-FERPlus, Pose-FERPlus, Occlusion-AffectNet, and Pose-AffectNet). These subsets are manually collected by reference [15]. There are six occlusion types for the occlusion test sets, namely wearing mask, wearing glasses, objects in bottom face, objects in upper face, objects in left/right face, and non-occlusion. In addition, the images in the pose variant test set were divided into two categories according to face with pitch or yaw angle greater than 30° or greater than 45°. Some samples are illustrated in Fig. 4.

4.2 Implementation details

We detected and aligned face regions using MTCNN for all images in the above-mentioned datasets and then resized them to $3 \times 224 \times 224$. Random cropping and horizontal flipping, etc. are further employed to avoid over-fitting. To make a fair comparison with the state-of-the-art methods, we adopt ResNet-18 [55] as a backbone, which is pre-trained on the MS-Celeb-1 M face recognition dataset [63]. In the GFE-AN stream, parameters λ_1 and λ_2 are 0.01. The dimension of high-level feature $d = C/F$ and the reduction ratio is 2 in the SA block. The split number F in the MFF-AN stream is 4. The details of these parameter settings can be found in Sects. 4.4.1, 4.4.3, and 4.4.6, respectively.

Our DSAN is implemented end-to-end with the Pytorch toolbox[1] on Ubuntu 18.04, 64 G RAM, 2 NVIDIA GeForce RTX 2080 Ti GPU. We train the first branch with $\mathcal{L}_{FR\_C}$ and $\mathcal{L}_{CCS}$ using Stochastic Gradient Descent (SGD). The initial learning rate and the momentum are 0.9, and the weight decay is 1e−4. Besides, we train the second branch and the two classifiers with $\mathcal{L}_{CE}$ and $\mathcal{L}_{FR\_CE}$ using Adam optimizer. We set the base learning rate as 1e−3 on RAF-DB and FERPlus and 1e−4 on AffectNet, exponentially decayed by a gamma of 0.6. The weight decay parameter is 1e−4 during the whole training phase. It is noteworthy that the whole model is jointly optimized with $\mathcal{L}_{FR}$ and $\mathcal{L}_{CE}$. Parameter α is used to balance the above two loss functions and empirically set as 0.4, and the parameter setting details can be seen in Sect. 4.4.4. We train the whole model with batch size 128 and stop training after 40 epochs. The number of Parameters and FLOPs of DSAN is 11.33 M and 1821.82 M, respectively.

4.3 Comparison with the state-of-the-arts

To further show the superiority of our method, we report the specific expression recognition accuracy on RAF-DB, FERPlus, AffectNet-7, and AffectNet-8 in Fig. 5. We also conduct several experiments on occlusion and pose variation datasets to examine our method in case of realistic scenarios.

[1] https://pytorch.org/.

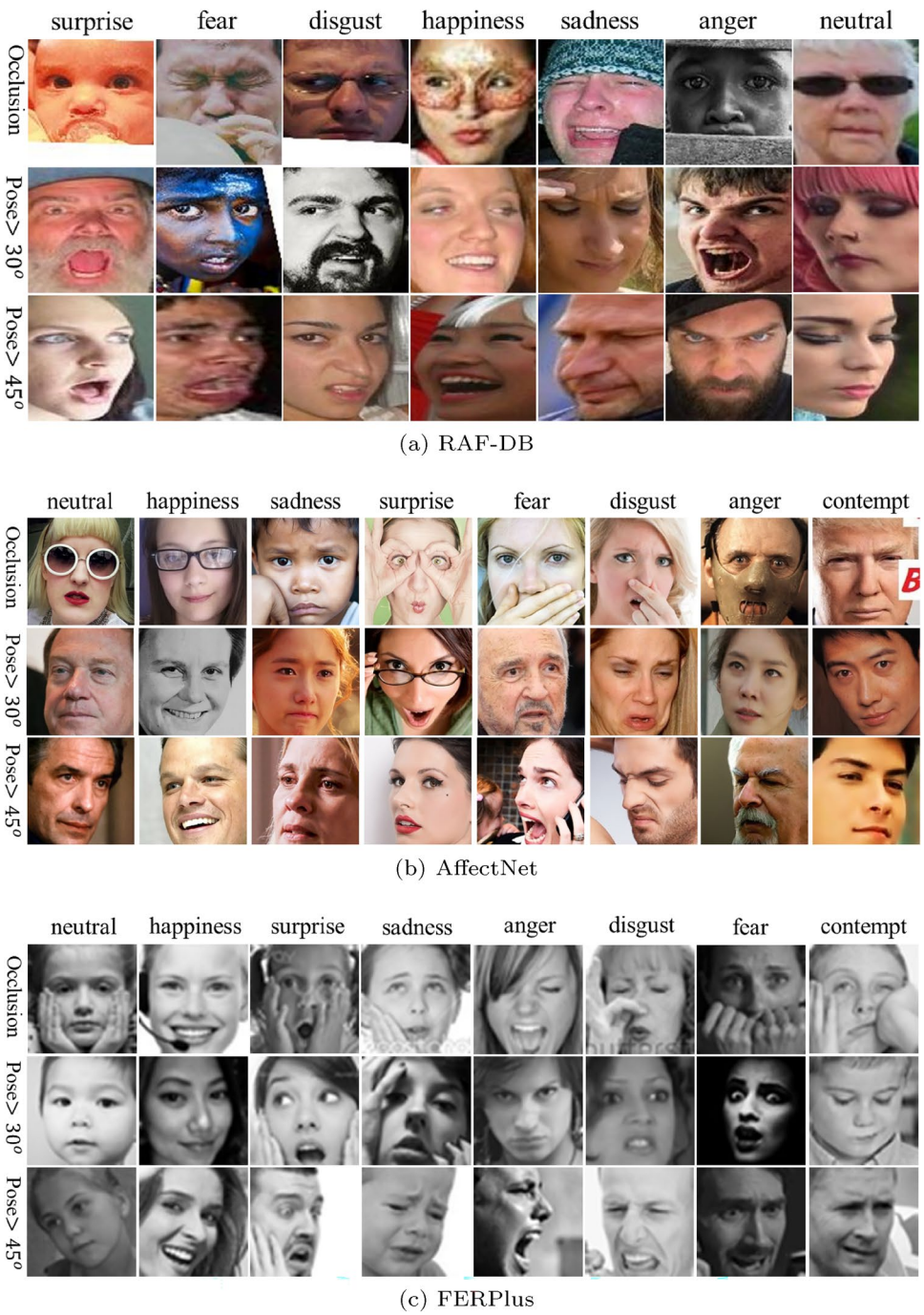
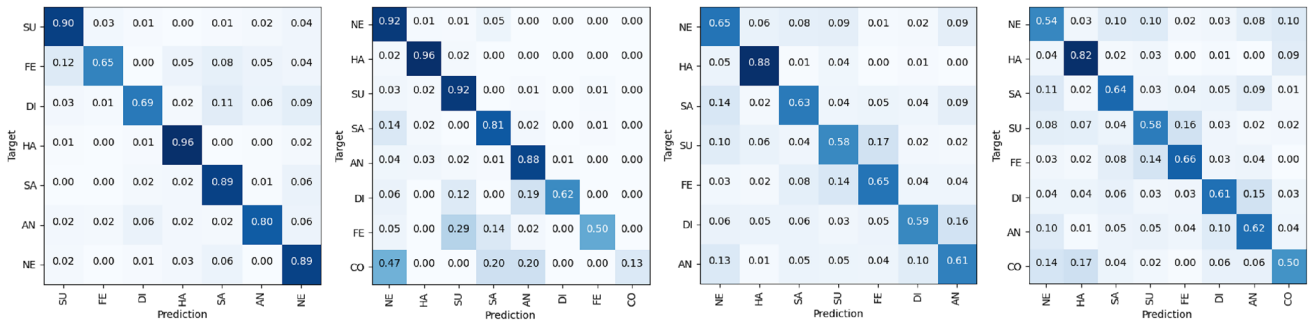


Fig. 4 Some examples of reference [15] collected occlusion and pose variant test subsets from RAF-DB, AffectNet, and FERPlus

4.3.1 Results on RAF-DB

To verify the effectiveness of DSAN, we added two metrics, the number of parameters and FLOPs, to measure the computing overhead of our model. Table 1 lists the experimental results. Among them, gACNN, RAN, and MA-Net tend to learn small weights for patches or local features of non-facial regions to address facial occlusion and pose variation. Although the above three methods have progressed in FER in the wild, they ignored parameter storage or computational efficiency and thus are difficult to transfer to mobile devices. DAN presents a multi-head cross attention network for FER and achieves the highest accuracy and average accuracy. Our DSAN utilizes a global feature element-based attention network to recalibrate the feature vector and a multiple feature fusion-based attention network to learn a broader feature representation, which gets the top accuracy value and the second-best average accuracy value. Importantly, our DSAN is more outstanding in Params and MFLOPs than other algorithms.



(a) RAF-DB (b) FERPlus (c) AffectNet-7 (d) AffectNet-8

Fig. 5 The confusion matrices of our DSAN on the test set for the RAF-DB, FER+, AffectNet-7, and AffectNet-8. Best viewed in color and zoom in (colour figure online)

Table 1 Comparisons with the state-of-the-art methods on RAF-DB

Method	Year	Backbone	Params (M)	MFLOPs	Acc.	Avg. Acc.
gACNN [14]	2019	VGG-16	>134.29	>4109.48	85.07	–
RAN [15]	2020	ResNet-18	11.19	11,192.32	86.90	79.57
MA-Net [17]	2020	ResNet-18	50.55	3737.6	88.40	–
DACL [47]	2021	ResNet-18	103.04	20,193.28	87.78	80.44
DMUE [64]	2021	ResNet-18	>11.18	-	88.76	–
DMUE [64]	2021	ResNet-50	>23.52	-	89.42	–
DAN [10]	2021	ResNet-18	19.72	2283.52	**89.70**	**85.32**
PACVT [16]	2022	ResNet-18	28.39	–	88.21	82.14
VTFF [20]	2022	ResNet-18	51.80	–	88.14	81.86
AR-TE-CATFFNet [21]	2023	ResNet-18	32.12	4207.62	89.50	–
Baseline	2023	ResNet-18	**11.18**	**1818.56**	86.93	79.14
DSAN	2023	ResNet-18	11.33	1821.82	**89.70**	82.57

The best results are shown in bold, and the second-best results are underlined

Table 2 Comparisons with the state-of-the-art methods on FERPlus

Method	Year	Backbone	Acc.
SeNet50 [65]	2018	SeNet50	88.80
RAN [15]	2019	ResNet-18/VGG16	88.55/89.16
SCN [22]	2020	ResNet-18/ResNet-50	88.01/89.35
MVT [66]	2021	Deit-S	89.22
DMUE [64]	2021	ResNet-18/ResNet-50	88.64/89.51
PACVT [16]	2022	ResNet-18	88.72
VTFF [20]	2022	ResNet-18	88.81
Baseline	2023	ResNet-18	86.96
DSAN	2023	ResNet-18	**89.93**

The best results are shown in bold, and the second-best results are underlined

Because it not only considers how to improve the accuracy of FER but also how to save the calculated parameters as much as possible when constructing the DSAN framework. For example, sparse attention in GFE-AN utilizes a two-layer structure to save model parameters, and MFF-AN divides the feature vector into multiple local features for parallel computation to achieve a lighter than an FC layer.

4.3.2 Results on FERPlus

We compare DSAN with several state-of-the-art methods on FERPlus in Table 2. RAN was designed for facial occlusion and pose variation, while SCN and DMUE are designed for annotation ambiguity. The latest method, named MVT, simultaneously addressed the above two problems and achieved an outstanding result. In addition, SCN and DMUE pre-trained their model on ResNet-50 to obtain gains of 1.34% and 0.87%. In contrast, our method achieves the best performance despite being only designed for non-facial regions and adopting the pre-trained ResNet-18 as the backbone.

Table 3 Comparison to the state-of-the-art results on AffectNet

Method	Year	Backbone	Acc.
(a) Results on AffectNet-7			
gACNN [14]	2019	VGG-16	58.78
RAN [15]	2020	ResNet-18	59.50
MA-Net [17]	2020	ResNet-18	64.53
DACL [47]	2021	ResNet-18	65.20
MVT [66]	2021	Deit-S	64.57
EfficientFace [67]	2021	shuffleNet-V2	63.70
DAN [10]	2021	ResNet-18	65.69
DENet [68]	2023	ResNet-50	59.74
AR-TE-CATFFNet [21]	2023	ResNet-18	65.66
Baseline	2023	ResNet-18	63.57
DSAN	2023	ResNet-18	**65.77**
(b) Results on AffectNet-8			
SCN [22]	2020	ResNet-18	60.23
MA-Net [17]	2020	ResNet-18	60.29
EfficientFace [67]	2021	ShuffleNet-V2	59.89
MVT [66]	2021	Deit-S	61.40
DMUE [64]	2021	ResNet-18/ResNet-50	62.84/**63.11**
DAN [10]	2021	ResNet-18	62.09
PACVT [16]	2022	ResNet-18	60.68
VTFF [20]	2022	ResNet-18	61.85
Baseline	2023	ResNet-18	58.52
DSAN	2023	ResNet-18	62.13

The best results are shown in bold, and the second-best results are underlined

Table 4 Comparison to the state-of-the-art results on occlusion and pose variant datasets

Method	Year	Occlusion	Pose (> 30)	Pose (> 45)
(a) Results on Occlusion-RAF-DB, Pose-RAF-DB				
ResNet-18 [15]	2016	80.19	84.04	83.15
RAN [15]	2020	82.72	86.74	85.20
MA-Net [17]	2020	83.65	87.89	87.99
EfficientFace [67]	2021	83.24	88.13	86.92
MVT [66]	2021	85.17	87.99	88.40
VTFF [20]	2022	83.95	87.97	88.35
DENet [68]	2023	**85.44**	85.57	86.92
DSAN	2023	85.31	**89.57**	**89.43**
(b) Results on Occlusion-FER+, Pose-FER+				
ResNet-18 [15]	2016	73.33	78.11	75.50
RAN [15]	2020	83.63	82.23	80.40
VTFF [20]	2022	84.79	88.29	87.20
DSAN	2023	**86.61**	**89.83**	**89.42**
(c) Results on Occlusion-AffectNet, Pose-AffectNet				
ResNet-18 [15]	2016	49.48	50.10	48.50
RAN [15]	2020	58.50	53.90	53.19
MA-Net [17]	2020	59.59	57.51	57.78
EfficientFace [67]	2021	59.88	57.36	56.87
VTFF [20]	2022	**62.98**	60.61	61.00
DENet [68]	2023	57.39	54.23	53.81
DSAN	2023	62.66	**60.80**	**61.32**

The best results are shown in bold, and the second-best results are underlined

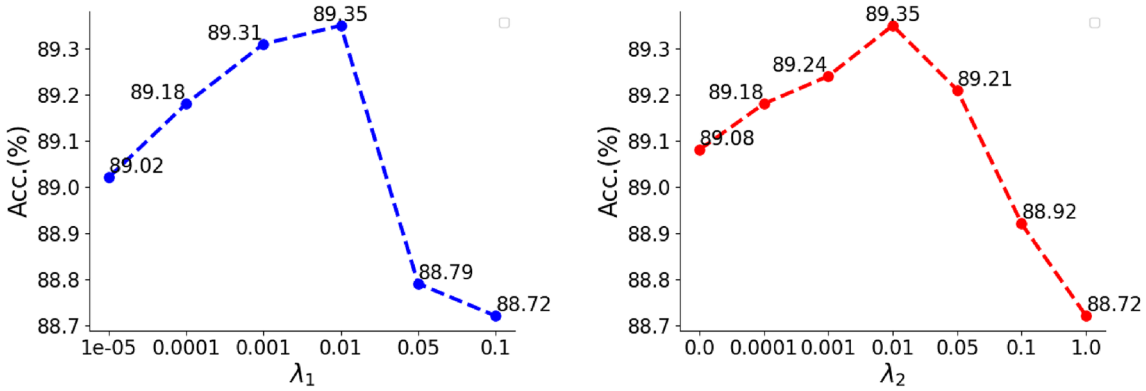
4.3.3 Results on AffectNet

AffectNet has an imbalanced training set but a balanced validation set. Consistent with RAN, SCN, and MA-Net, we employed an oversampling strategy in our experiments to overcome this issue. Table 3 presents the comparison results on AffectNet-7 and AffectNet-8. We can observe that DSAN obtains the highest FER accuracy of 65.77% on AffectNet with 7 expression categories. We obtain a comparable result (third highest accuracy) on AffectNet with 8 expression categories. Results of MA-Net, MVT, EfficientFace, DAN, and our DSAN all show a larger gap between the accuracy of AffectNet-7 and AffectNet-8. Moreover, as shown in Fig. 5c, d, the recognition accuracy of each category in AffectNet-8 showed an overall downward trend compared to AffectNet-7, and the results on *Contempt* were lowest on both datasets. That is because the category distribution of AffectNet is more unbalanced after adding the eighth category samples, and there exist lots of noise labels in the *Contempt* samples.

4.3.4 Results on occlusion and pose variant datasets

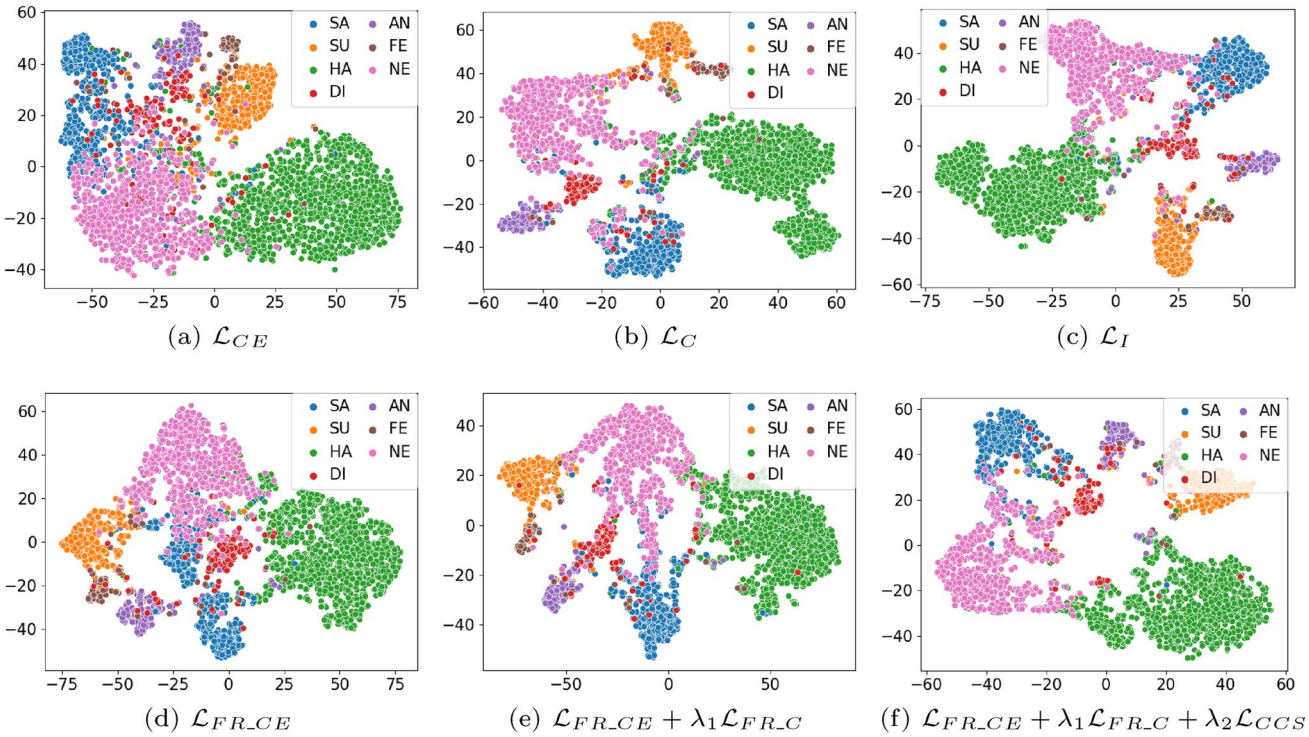
To verify the robustness of our method to occlusion and variant head pose, we conduct several experiments on occlusion-RAF-DB, pose-RAF-DB, occlusion-FERPlus, pose-FERPlus, occlusion-AffectNet, and pose-AffectNet. Table 4 shows the experiment results of DSAN and its comparison methods under corresponding subsets.

Our DSAN is specially designed for occlusion and pose variation. As shown in Table 4(a) and (b), our method obtains the highest accuracy and outperforms the state-of-the-art methods with a large margin in each case. Besides, DSAN achieves promising results on Occlusion-AffectNet and Pose-AffectNet, as shown in Table 4(c). It should be noted that our method performs more robustly to occlusion and poses variation on RAF-DB and FERPlus. That may be due to AffectNet containing too many noise labels, impeding the model's robust feature representation. Overall, these results verify that our DSAN has outstanding robustness towards occlusion and pose variation.



(a) The accuracy with different λ_1. (b) The accuracy with different λ_2.

Fig. 6 Ablation studies for the different values of the λ_1 and λ_2 in $\mathcal{L}_{FR}$ on RAF-DB (λ_1 and λ_2 represent the balance parameters for $\mathcal{L}_{FR\_C}$ and $\mathcal{L}_{CCS}$). Larger values correspond to better performance



(a) $\mathcal{L}_{CE}$ (b) $\mathcal{L}_C$ (c) $\mathcal{L}_I$

(d) $\mathcal{L}_{FR\_CE}$ (e) $\mathcal{L}_{FR\_CE} + \lambda_1\mathcal{L}_{FR\_C}$ (f) $\mathcal{L}_{FR\_CE} + \lambda_1\mathcal{L}_{FR\_C} + \lambda_2\mathcal{L}_{CCS}$

Fig. 7 The visualization study of the feature distribution obtained by GFE-AN under the supervision of **a** $\mathcal{L}_{CE}$, **b** $\mathcal{L}_C$, **c** $\mathcal{L}_I$, **d** $\mathcal{L}_{FR\_CE}$, **e** $\mathcal{L}_{FR\_CE} + \lambda_1\mathcal{L}_{FR\_C}$, and **f** $\mathcal{L}_{FR\_CE} + \lambda_1\mathcal{L}_{FR\_C} + \lambda_2\mathcal{L}_{CCS}$ on RAF-DB testset. Where $\mathcal{L}_{CE}$, $\mathcal{L}_C$, $\mathcal{L}_I$, $\mathcal{L}_{FR\_CE}$, $\mathcal{L}_{FR\_C}$, and $\mathcal{L}_{CCS}$ represent cross-entropy loss, center loss, island loss, feature recalibration based cross-entropy loss, feature recalibration based center loss, and class center separation, respectively. The dots with different colors represent different facial emotions, best viewed in color and zoom-in (colour figure online)

4.4 Ablation studies

As shown in Fig. 2, our DSAN mainly consists of GFE-AN and MFF-AN. To better understand and show the effectiveness of our method, we conduct quantitative and qualitative analyses of the critical parameters and components of these two branches on the RAF-DB dataset. In addition, we evaluate the influence of sparse attention structure on FER.

Table 5 Evaluation of the main components of $\mathcal{L}_{FR}$ and their similar loss functions in terms of accuracy (%) on RAF-DB

$\mathcal{L}_{CE}$	$\mathcal{L}_{C}$	$\mathcal{L}_{I}$	$\mathcal{L}_{FR}$			Acc.
			$\mathcal{L}_{FR\_CE}$	$\mathcal{L}_{FR\_C}$	$\mathcal{L}_{CCS}$	
✓	×	×	×	×	×	88.53
×	✓	×	×	×	×	88.85
×	×	✓	×	×	×	88.96
×	×	×	✓	×	×	88.72
×	×	×	✓	✓	×	89.18
×	×	×	✓	✓	✓	**89.35**

The best results are shown in bold

4.4.1 Visualization of the λ_1 and λ_2 in $\mathcal{L}_{FR}$

λ_1 and λ_2 represent the contribution degree of $\mathcal{L}_{FR\_C}$ and $\mathcal{L}_{CCS}$ to facial expression recognition. We explore the impacts of the different values of λ_1 and λ_2 on the recognition performance of GFE-AN. The experiment results are given in Fig. 6. We first fix $\lambda_2 = 0.01$ and set the value of λ_1 from 1e−05 to 0.1. As shown in Fig. 6a, The recognition accuracy increases gradually when λ_1 is in the interval from 1e−05 to 0.01 and decreases sharply when λ_1 is greater than 0.01. This indicates that $\mathcal{L}_{FR\_C}$ helps improve the recognition accuracy of DSAN. Subsequently, we fix $\lambda_1 = 0.01$ and set the value of λ_2 from 0.0 to 1.0 to test the importance of $\mathcal{L}_{CCS}$. As shown in Fig. 6b, when $\lambda_2 = 0.0$, the performance of GFE-AN without $\mathcal{L}_{CCS}$ decreases. In addition, increasing λ_2 means expanding the distance between class centers. It is not desirable to separate the class centers as much as possible because there are still some hard samples whose features are not in their cluster. GFE-AN achieves the top performance when the value of λ_2 is set to 0.01. In the following, λ_1 and λ_2 are set to 0.01 by default.

4.4.2 Evaluation of different loss functions

To show the effectiveness of our proposed loss functions, we visualize the feature distributions of samples from the test subset of RAF-DB by using t-SNE [69]. As shown in Fig. 7, a total of 3068 images from 7 classes were used for this experiment. We give the changing processes of feature distribution of GFE-AN under the influence of the three components in the feature recalibration loss function. $\mathcal{L}_{FR\_CE}$ is refactored based on $\mathcal{L}_{CE}$, which urges the model to remove feature elements that interfere with classification. To make intra-class aggregation and inter-class separation, we added $\mathcal{L}_{FR\_C}$ and $\mathcal{L}_{CCS}$ loss functions for supervision training. As described in Sect. 3.2, $\mathcal{L}_{FR}$ can reduce to cross-entropy loss, center loss, and island loss under certain conditions. We compare (b) with (e) and (c) with (f), respectively, in Fig. 7. The observations show that our constructed loss functions help extract robust feature representation. For example, the last two figures show that the number of *Disgust* samples (red dots) in the *Happy* cluster (green dots) has decreased. To further demonstrate the effectiveness of each component in $\mathcal{L}_{FR}$, we design an ablation study to investigate the above proposed six loss functions on the RAF-DB dataset. Table 5 shows the experiment results, and we conclude several observations. First, the recognition accuracy of $\mathcal{L}_{FR\_CE}$ is higher than baseline but performs worse than the naive center loss and island loss. The reason is that the feature recalibration process removes many feature elements that interfere with classification, which causes the intra-class features to diverge, but the inter-class features still overlap. Second, we obtain the most remarkable improvement by adding a $\mathcal{L}_{FR\_C}$, which improves the accuracy of the $\mathcal{L}_{FR\_CE}$ from 88.72% to 89.18%. We can infer that $\mathcal{L}_{FR\_C}$ is the most contributed loss function for $\mathcal{L}_{FR}$. Third, adding the $\mathcal{L}_{CCS}$ can further boost $\mathcal{L}_{FR}$ by 0.17%, which achieves the best performance.

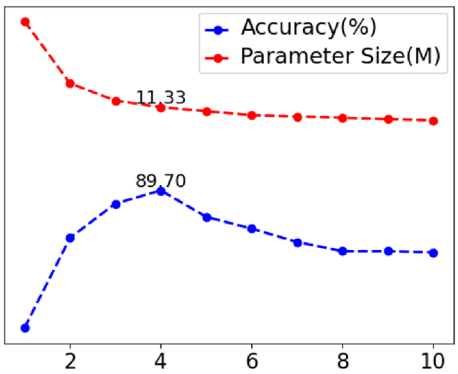


Fig. 8 Ablation studies for the various values of F in MFF-AN on RAF-DB. The X-axis represents the number of local features F, and the Y-axis represents the accuracy and parameter size of DSAN

4.4.3 Effectiveness of *F* for FER

The number of local features in MFF-AN affects the performance and size of the whole model. Fig. 8 shows the

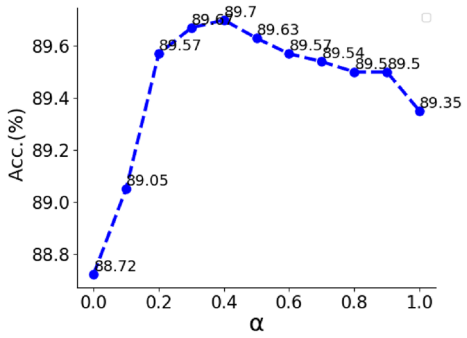


Fig. 9 Evaluation of different α on RAF-DB

Table 6 Evaluation of the main components of DSAN in terms of accuracy (%) on RAF-DB

GFE-AN		MFF-AN		Acc.
HFE_1	SA	HFE_2	LFA	
×	×	×	×	87.68
✓	×	×	×	88.53
✓	✓	×	×	89.35
×	×	✓	×	88.56
×	×	✓	✓	88.72
✓	×	✓	×	88.92
✓	×	✓	✓	89.18
✓	✓	✓	×	89.54
✓	✓	✓	✓	**89.70**

The best results are shown in bold

HFE_1 and HFE_2 represent the high-level feature extractor in the first stream GFE-AN and the second stream MFF-AN, respectively. SA is sparse attention, and LFA is local feature attention

accuracy results and parameter sizes with changing the number of local features on RAF-DB. As the increasing of F, the number of parameters of DSAN drops sharply then levels off gradually, and the accuracy of DSAN first increases then decreases. To balance our model's parameter size and classification accuracy, F defaults to 4 in the following.

4.4.4 Evaluation of different α

α is a parameter to balance the loss function and prediction probability between GFE-AN and MFF-AN. To explore the impact of α on our DSAN, we evaluate it from 0.0 to 1.0. Fig. 9 shows the results on the RAF-DB dataset. Especially when α is 0.0, our DSAN is reduced to MFF-AN, the accuracy of which reaches 88.72%. Furthermore, when α is 1.0, our DSAN can be regarded as GFE-AN, and the accuracy of

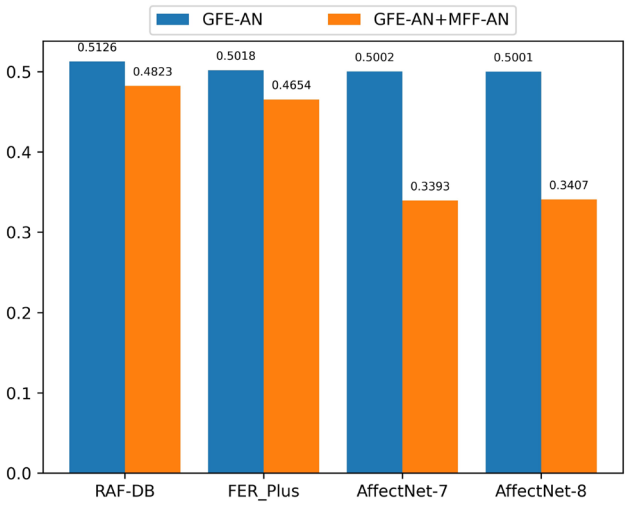


Fig. 10 The average probability of $a_{i,j} = 0$ in GFE-AN and GFE-AN+MFF-AN

which reaches 89.35%. Moreover, the default $\alpha = 0.4$ obtains the best performance. Therefore, we may infer that the MFF-AN is essential to complement the lost facial expression information of GFE-AN.

4.4.5 Component analysis of DSAN

To evaluate the effectiveness of the critical modules in DSAN, we conduct ablation studies for GFE-AN and MFF-AN on RAF-DB. Specifically, GFE-AN contains a high-level feature extractor HFE_1 and sparse attention block SA. MFF-AN comprises multiple high-level feature extractors HFE_2 and local feature attention blocks LFA. The experimental results as shown in Table 6.

The ResNet-18 with max-pooling is pre-trained on the MS-Celeb-1 M face recognition dataset and employed as a baseline (first row). We can see that after adding HFE_1 and HFE_2 to the baseline, the results improved by 0.85% and 0.88% in GFE-AN and MFF-AN. It suggests that higher-level features have more excellent recognition capability. In addition, employing SA and LFA further boosts the performance of DSAN, and SA performs better than LFA. The attention weights in SA are learned under the joint supervision of multiple loss functions, so the weighted features can quickly achieve intra-class aggregation and inter-class separation. At the same time, LFA performs a simple weighted summation for local features to obtain the optimal feature representation. Moreover, when adding another branch, the GFE-AN and MFF-AN are improved from 89.35% and 88.72% to 89.70%, respectively. This is because GFE-AN can effectively alleviate the interference of non-facial regions in natural scenes, while MFF-AN provides more complete facial expression information.

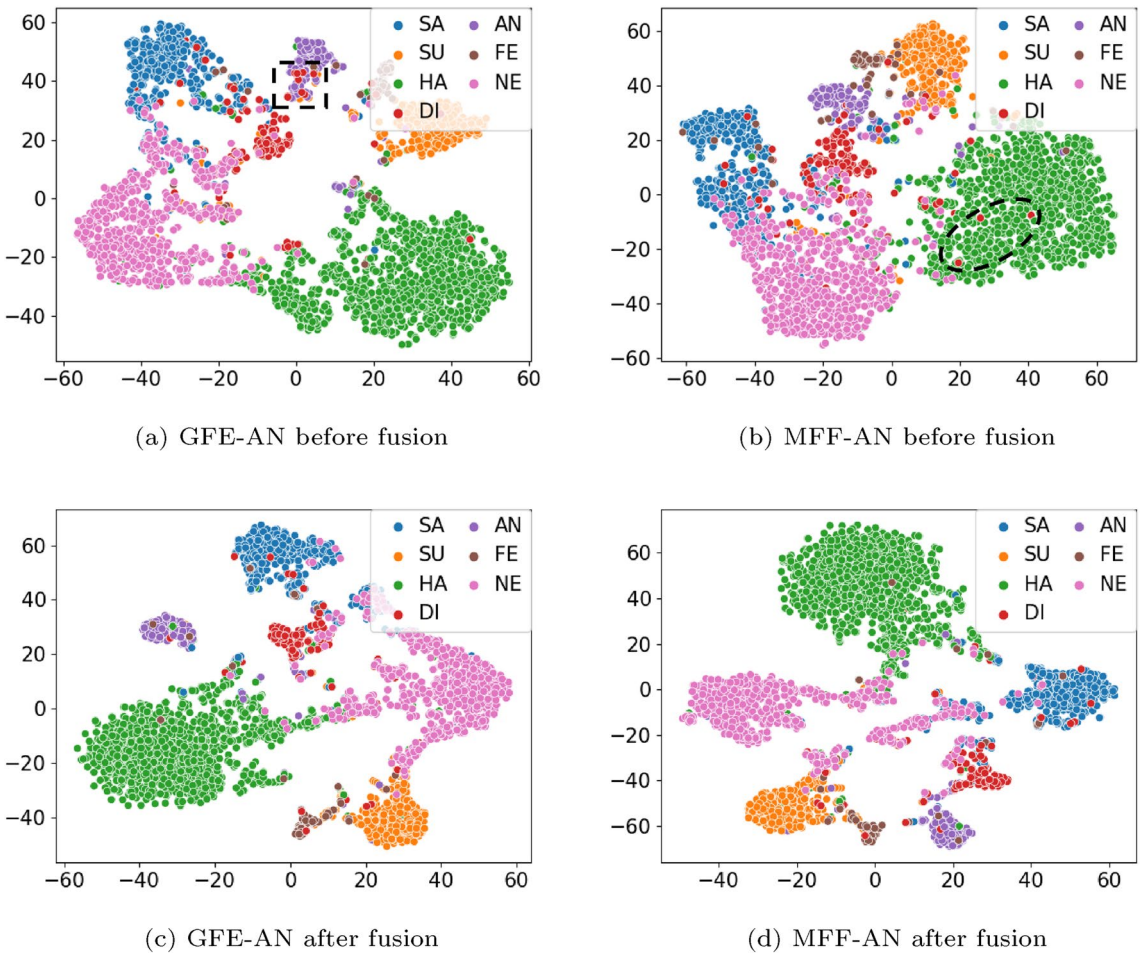


Fig. 11 Visualization study of the feature distribution for each module of DSAN before and after fusion. **a** GFE-AN before fusion; **b** MFF-AN before fusion; **c** GFE-AN after fusion; **d** MFF-AN after fusion. The dots with different colors represent different facial emotions, best viewed in color and zoom-in (colour figure online)

We calculate the average value of $a_{i,j} = 0$ (expressed as $P_{a_{i,j}=0}$) to represent the proportion of the removed feature elements in a sample on RAF-DB, FERPlus, and AffectNet. The experimental results are shown in Fig. 10. It is clear that $P_{a_{i,j}=0} > 0.5$ when only using the first branch GFE-AN, which means that GFE-AN removes the feature elements that interfere with the classification of expressions while the information about facial expressions also is suppressed. When the second branch MFF-AN is added, the value of $P_{a_{i,j}=0}$ decreases obviously, which means that the second branch can effectively supplement the lost expression information of GFE-AN.

To further confirm the above view, we use t-SNE [69] to visualize the characteristics of the two modules of DSAN before and after fusion. As shown in Fig. 11a, the GFE-AN branch can reduce noise information to an extent, but the extracted facial expression information is insufficient for differentiating similar expressions, making classification challenging. For instance, some *Disgust* samples in the *Anger* cluster, marked by a black rectangle, are difficult to distinguish. In Fig. 11b, the features extracted by MFF-AN contain redundant information because natural scene samples contain more noise.

Table 7 Accuracy and parameter size for sparse attention block at different reduction ratios on RAF-DB

Layer	r	Params	Acc
1	1	16,768	89.28
2	1	33,536	89.11
	2	16,960	**89.35**
	3	11,262	<u>89.34</u>
	4	8672	89.31
	5	6859	89.24
	6	5823	89.11

The best results are shown in bold, and the second-best results are underlined

It leads to confusion between positive affective samples (*Happy*) and negative affective samples (including *Disgust*, *Sadness*, *Anger*, and *Fear*). For example, the red dots *Disgust* marked with a black ellipse are easily misclassified as *Happy*.

When GFE-AN and MFF-AN are fused, we observed two distinct changes. Firstly, in Fig. 11c, the *Disgust* samples in the *Anger* cluster are significantly reduced, indicating that MFF-AN can provide rich facial expression information for GFE-AN. Secondly, in Fig. 11d, the *Disgust* samples remarked with black ellipse were removed from the *Happy* feature cluster, and the feature map of MFF-AN showed the phenomenon of intra-class compact inter-class separation like GFE-AN. It further verifies that MFF-AN is affected by GFE-AN. In addition, the fused MFF-AN eliminates facial noise while retaining rich expression information.

4.4.6 Structure analysis of sparse Attention

To investigate the trade-off between performance and computing overhead, we explore the impact of r on the first stream. As shown in Table 7, We conduct experiments by forming a bottleneck with two FC layers. The input and output dimensions of SA are d, and the output dimension of the first FC layer is d/r . We can observe that when the number of FC layers is 2, the number of parameters decreases as the reduction ratio increases. Concurrently, the accuracy of GFE-AN achieves the top value when $r = 2$. Moreover, GFE-AN achieves the best performance when using double-layer FC, and its parameter number is comparable to those of single-layer FC. Therefore, the number of layers and the reduction ratio is set to 2.

5 Discussion

In this paper, our DSAN achieved outstanding results with few parameters and FLOPs in the FER task on all three ITW datasets. Compared to the existing spatial attention methods based on mid-level local features, our proposed DSAN focused on the elements of the high-level feature vector. As shown in Fig. 7, sparse attention training with feature recalibration loss assigns different weights to each element, achieving inter-class aggregation and inter-class separation. This can effectively cope with the impact of facial occlusion and pose variations, but some facial expression information would inevitably be lost. Therefore, another branch is proposed to extract rich expression features to compensate for the deficiency.

Despite the significant improvements presented in our study, some limitations warrant further research. First, the number of elements with 0 weight in sparse attention is uncontrollable. Specifically, GFE-AN can focus on significant facial expression feature elements automatically, and some non-facial feature elements have also been suppressed when the weight is zero to improve the model's facial expression recognition ability effectively. But while using sparse attention to suppress non-facial elements, some facial elements are also forced to be ignored. Hence, only training the GFE-AN branch will result in losing information about facial expressions. If a mechanism could automatically control the number of $a_{i,j} = 0$, the non-facial problem in the wild can be solved, while the expression information could be preserved as much as possible. This question will be left for our future work. Second, our DSAN is not light enough. The most significant contribution of our model is to improve FER performance in the wild, thus ignoring the saving of model parameters and the acceleration of calculation time. Achieving a lighter model requires replacing the existing backbones, such as ResNet-18 and VGG16, and the study of FER lightweight models in uncontrolled environments remains a challenge.

6 Conclusion

This paper proposes an end-to-end network named Dual Stream Attention Network (DSAN) to address occlusion and pose variation for facial expression recognition in the wild. Precisely, DSAN mainly consists of two parallel networks: GFE-AN and MFF-AN. GFE-AN designs a feature recalibration loss function, which forces the network to selectively emphasize the feature elements significant to expression classification and suppress those unrelated to expression. MFF-AN adaptively fuses multiple local features with broader representational capabilities, further improving the performance of DSAN. Extensive experiments on three popular ITW datasets, including RAF-DB, FERPLus, and AffectNet, show that the proposed DSAN achieves state-of-the-art results with relatively little storage space and computing time. To verify the superiority and effectiveness of DSAN to occlusion and pose variation, we test our model on several realistic occlusion and pose variation subsets. The results demonstrate that the proposed approach is robust and outperforms the most recent ITW FER methods.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grants (Grant Nos 61802184, 61972204, 62103110, 62103192, and 62102002).

Author contributions Conceptualization, Methodology, Data curation, Validation, Visualization, and Writing - original draft preparation: Hui Tang; Funding acquisition and Resources: Zhong Jin; Supervision: Yichang Li and Zhong Jin; Formal analysis, Investigation, and Writing - review and editing: Hui Tang, Yichang Li, and Zhong Jin.

Data availability and access All datasets supporting the findings of this study are available and have been described in Sect. 4.1.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article.

References

1. Darwin C (1872) The expression of the emotions in man and animals. John Murray, London
2. Tian Y-I, Kanade T, Cohn JF (2001) Recognizing action units for facial expression analysis. IEEE Trans Pattern Anal Mach Intell 23(2):97–115
3. Li S, Deng W (2020) Deep facial expression recognition: a survey. IEEE Trans Affect Comput 13(3):1195–1215
4. Benitez-Quiroz CF, Srinivasan R, Martinez AM (2018) Discriminant functional learning of color features for the recognition of facial action units and their intensities. IEEE Trans Pattern Anal Mach Intell 41(12):2835–2845
5. Kanade T, Cohn JF, Tian Y (2000) Comprehensive database for facial expression analysis. In: Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), pp 46–53
6. Shih FY, Chuang C-F, Wang PS (2008) Performance comparisons of facial expression recognition in Jaffe database. Int J Pattern Recognit Artif Intell 22(3):445–459
7. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.-Workshops (CVPRW), pp 94–101
8. Zhao G, Huang X, Taini M, Li SZ, PietikäInen M (2011) Facial expression recognition from near-infrared videos. Image Vis Comput 29(9):607–619
9. Valstar M, Pantic M (2010) Induced disgust, happiness and surprise: an addition to the MMI facial expression database. In: Proc. 3rd Int. Workshop Emotion (satell. of LREC): Corpora Res. Emotion Affect. Paris, France
10. Wen Z, Lin W, Wang T, Xu G (2021) Distract your attention: multi-head cross attention network for facial expression recognition. arXiv prepr. arXiv:2109.07270
11. Li S, Deng W, Du J (2017) Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 2852–2861
12. Barsoum E, Zhang C, Ferrer CC, Zhang Z (2016) Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proc. 18th ACM Int. Conf. Multimodal Interact., pp 279–283
13. Mollahosseini A, Hasani B, Mahoor MH (2017) Affectnet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans Affect Comput 10(1):18–31
14. Li Y, Zeng J, Shan S, Chen X (2018) Occlusion aware facial expression recognition using CNN with attention mechanism. IEEE Trans Image Process 28(5):2439–2450
15. Wang K, Peng X, Yang J, Meng D, Qiao Y (2020) Region attention networks for pose and occlusion robust facial expression recognition. IEEE Trans Image Process 29:4057–4069
16. Liu C, Hirota K, Dai Y (2023) Patch attention convolutional vision transformer for facial expression recognition with occlusion. Inf Sci 619:781–794
17. Zhao Z, Liu Q, Wang S (2021) Learning deep global multi-scale and local attention features for facial expression recognition in the wild. IEEE Trans Image Process 30:6544–6556
18. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: Proc. Eur. Conf. Comput. Vis. Springer, pp 818–833
19. Ruan D, Yan Y, Lai S, Chai Z, Shen C, Wang H (2021) Feature decomposition and reconstruction learning for effective facial expression recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 7660–7669
20. Ma F, Sun B, Li S (2023) Facial expression recognition with visual transformers and attentional selective fusion. IEEE Trans Affect Comput 14(2):1236–1248
21. Sun M, Cui W, Zhang Y, Yu S, Liao X, Hu B, Li Y (2023) Attention-rectified and texture-enhanced cross-attention transformer feature fusion network for facial expression recognition. IEEE Trans Ind Inform 19:11823–11832
22. Wang K, Peng X, Yang J, Lu S, Qiao Y (2020) Suppressing uncertainties for large-scale facial expression recognition. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp 6897–6906
23. Li H, Wang N, Ding X, Yang X, Gao X (2021) Adaptively learning facial expression representation via cf labels and distillation. IEEE Trans Image Process 30:2016–2028
24. Jin X, Lai Z, Jin Z (2021) Learning dynamic relationships for facial expression recognition based on graph convolutional network. IEEE Trans Image Process 30:7143–7155
25. Ding H, Zhou SK, Chellappa R (2017) FaceNet2ExpNet: regularizing a deep face recognition net for expression recognition. In: Proc. 2017 12th IEEE Int. Conf. Aut. Face Gesture Recognit. (FG 2017), pp 118–126
26. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C (2018) MobileNetV2: inverted residuals and linear bottlenecks. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 4510–4520
27. Fu Y, Wu X, Li X, Pan Z, Luo D (2020) Semantic neighborhood-aware deep facial expression recognition. IEEE Trans Image Process 29:6535–6548
28. Nan Y, Ju J, Hua Q, Zhang H, Wang B (2022) A-MobileNet: an approach of facial expression recognition. Alex Eng J 61(6):4435–4444
29. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Proc. Neural Inf. Proces. Syst., pp 5999–6009
30. Happy S, George A, Routray A (2012) A real time facial expression classification system using local binary patterns. In: Proc. 2012 4th Int. Conf. Intell. Hum. Comput. Interact. IEEE, pp 1–5
31. Kaya H, Gürpinar, F, Afshar S, Salah AA (2015) Contrasting and combining least squares based learners for emotion recognition in the wild. In: Proc. ACM Int. Conf. Multimodal Interact., pp 459–466
32. Valstar MF, Mehu M, Jiang B, Pantic M, Scherer K (2012) Meta-analysis of the first facial expression recognition challenge. IEEE Trans Syst Man Cybern Part B (Cybernetics) 42(4):966–979
33. Kotsia I, Buciu I, Pitas I (2008) An analysis of facial expression recognition under partial facial image occlusion. Image Vis Comput 26(7):1052–1067
34. Bartlett MS, Littlewort G, Frank M, Lainscsek C, Fasel I, Movellan J (2005) Recognizing facial expression: machine learning and application to spontaneous behavior. In: Proc. IEEE Comput. Society Conf. Comput. Vis. Pattern Recognit., vol 2. IEEE, pp 568–573
35. Jiang B, Martinez B, Valstar MF, Pantic M (2014) Decision level fusion of domain specific regions for facial action recognition. In: Proc. 22th Int. Conf. Pattern Recognit. IEEE, pp 1776–1781
36. Berretti S, Del Bimbo A, Pala P, Amor BB, Daoudi M (2010) A set of selected sift features for 3D facial expression recognition. In: Proc. 20th Int. Conf. Pattern Recognit. IEEE, pp 4125–4128

37. Gritti T, Shan C, Jeanne V, Braspenning R (2008) Local features based facial expression recognition with face registration errors. In: Proc. 8th IEEE Int. Conf. Automatic Face Gesture Recognit. IEEE, pp 1–8
38. Umer S, Rout RK, Pero C, Nappi M (2022) Facial expression recognition with trade-offs between data augmentation and deep learning features. J Ambient Intell Humaniz Comput 13(2):721–735
39. Zhao X, Liang X, Liu L, Li T, Han Y, Vasconcelos N, Yan S (2016) Peak-piloted deep network for facial expression recognition. In: Proc. Eur. Conf. Comput. Vis.. Springer, pp 425–442
40. Wang Z, Chen J, Hoi SC (2021) Deep learning for image super-resolution: a survey. IEEE Trans Pattern Anal Mach Intell 43(10):3365–3387
41. Shao J, Cheng Q (2021) E-FCNN for tiny facial expression recognition. Appl Intell 51(1):549–559
42. Vo T-H, Lee G-S, Yang H-J, Kim S-H (2020) Pyramid with super resolution for in-the-wild facial expression recognition. IEEE Access 8:131988–132001
43. Lim B, Son S, Kim H, Nah S, Mu Lee K (2017) Enhanced deep residual networks for single image super-resolution. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit-Workshops., pp 136–144
44. Zeng J, Shan S, Chen X (2018) Facial expression recognition with inconsistently annotated datasets. In: Proc. Eur. Conf. Comput. Vis., pp 222–237
45. Wen Y, Zhang K, Li Z, Qiao Y (2016) A discriminative feature learning approach for deep face recognition. In: Proc. Eur. Conf. Comput. Vis. Springer, pp 499–515
46. Cai J, Meng Z, Khan AS, Li Z, O'Reilly J, Tong Y (2018) Island loss for learning discriminative features in facial expression recognition. In: Proc. 13th IEEE Int. Conf. Automatic Face Gesture Recognit.. IEEE, pp 302–309
47. Farzaneh AH, Qi X (2021) Facial expression recognition in the wild via deep attentive center loss. In: Proc. IEEE Winter Conf. Appl. Comput. Vis., pp 2402–2411
48. Li Y, Lu Y, Li J, Lu G (2019) Separate loss for basic and compound facial expression recognition in the wild. In: Proc. Asian Conf. Mach. Learn.. PMLR, pp 897–911
49. Mnih V, Heess N, Graves A (2014) Recurrent models of visual attention. In: Proc. Adv. Neural Inf. Process. Syst., pp 2204–2212
50. Chen W, Zhang D, Li M, Lee D-J (2020) STCAM: spatial-temporal and channel attention module for dynamic facial expression recognition. IEEE Trans Affect Comput 14:800–810
51. Hasani B, Mahoor MH (2017) Facial expression recognition using enhanced deep 3D convolutional neural networks. In: Proc. 2017 IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, pp 2278–2288. https://doi.org/10.1109/CVPRW.2017.282
52. Wu R, Zhang G, Lu S, Chen T (2020) Cascade EF-GAN: progressive facial expression editing with local focuses. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp 5021–5030
53. Ni R, Yang B, Zhou X, Cangelosi A, Liu X (2022) Facial expression recognition through cross-modality attention fusion. IEEE Trans Cogn Dev Syst 15:175–185
54. Gera D, Balasubramanian S (2021) Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition. Pattern Recognit Lett 145:58–66
55. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp 770–778
56. Zhou S-Y, Su C-Y (2021) A novel lightweight convolutional neural network, exquisitenetv2. arXiv prepr. arXiv:2105.09008
57. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp 7132–7141
58. Ma J, Ma Y, Li C (2019) Infrared and visible image fusion methods and applications: a survey. Inf Fusion 45:153–178
59. Zeng Z, Pantic M, Roisman GI, Huang TS (2008) A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell 31(1):39–58
60. Li S, Kang X, Fang L, Hu J, Yin H (2017) Pixel-level image fusion: a survey of the state of the art. Inf Fusion 33:100–112
61. Cai J, Meng Z, Khan AS, Li Z, O'Reilly J, Han S, Liu P, Chen M, Tong Y (2019) Feature-level and model-level audiovisual fusion for emotion recognition in the wild. In: Proc. IEEE Conf. Multimedia Inf. Process. Retr.. IEEE, pp 443–448
62. Goodfellow IJ, Erhan D, Carrier PL, Courville A, Mirza M, Hamner B, Cukierski W, Tang Y, Thaler D, Lee D-H (2013) Challenges in representation learning: a report on three machine learning contests. In: Proc. Int. Conf. Neural Inf. Process. Springer, pp 117–124
63. Guo Y, Zhang L, Hu Y, He X, Gao J (2016) MS-Celeb-1M: a dataset and benchmark for large-scale face recognition. In: Proc. Eur. Conf. Comput. Vis.. Springer, pp 87–102
64. She J, Hu Y, Shi H, Wang J, Shen Q, Mei T (2021) Dive into ambiguity: latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 6248–6257
65. Albanie S, Nagrani A, Vedaldi A, Zisserman A (2018) Emotion recognition in speech using cross-modal transfer in the wild. In: Proc. 26th ACM Int. Conf. Multimedia, pp 292–301
66. Li H, Sui M, Zhao F, Zha Z, Wu F (2021) MVT: mask vision transformer for facial expression recognition in the wild. arXiv prepr. arXiv:2106.04520
67. Zhao Z, Liu Q, Zhou F (2021) Robust lightweight facial expression recognition network with label distribution training. In: Proc. AAAI Conf. Artif. Intell., vol. 35, pp 3510–3519
68. Li H, Wang N, Yang X, Wang X, Gao X (2023) Unconstrained facial expression recognition with no-reference de-elements learning. IEEE Trans Affect Comput 15:173–185
69. Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9(11):2579–2605