future internet

MDPI

Article

Facial Expression Recognition Using Dual Path Feature Fusion and Stacked Attention

Hongtao Zhu [1,*], Huahu Xu [1,2], Xiaojin Ma [1] and Minjie Bian [1,2]

[1] School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
[2] Shanghai Shangda Hairun Information System Co., Ltd., Shanghai 200072, China
[*] Correspondence: keyboy@shu.edu.cn

Abstract: Facial Expression Recognition (FER) can achieve an understanding of the emotional changes of a specific target group. The relatively small dataset related to facial expression recognition and the lack of a high accuracy of expression recognition are both a challenge for researchers. In recent years, with the rapid development of computer technology, especially the great progress of deep learning, more and more convolutional neural networks have been developed for FER research. Most of the convolutional neural performances are not good enough when dealing with the problems of overfitting from too-small datasets and noise, due to expression-independent intra-class differences. In this paper, we propose a Dual Path Stacked Attention Network (DPSAN) to better cope with the above challenges. Firstly, the features of key regions in faces are extracted using segmentation, and irrelevant regions are ignored, which effectively suppresses intra-class differences. Secondly, by providing the global image and segmented local image regions as training data for the integrated dual path model, the overfitting problem of the deep network due to a lack of data can be effectively mitigated. Finally, this paper also designs a stacked attention module to weight the fused feature maps according to the importance of each part for expression recognition. For the cropping scheme, this paper chooses to adopt a cropping method based on the fixed four regions of the face image, to segment out the key image regions and to ignore the irrelevant regions, so as to improve the efficiency of the algorithm computation. The experimental results on the public datasets, CK+ and FERPLUS, demonstrate the effectiveness of DPSAN, and its accuracy reaches the level of current state-of-the-art methods on both CK+ and FERPLUS, with 93.2% and 87.63% accuracy on the CK+ dataset and FERPLUS dataset, respectively.

Keywords: deep learning; attention; facial expression recognition

check for updates

Citation: Zhu, H.; Xu, H.; Ma, X.; Bian, M. Facial Expression Recognition Using Dual Path Feature Fusion and Stacked Attention. *Future Internet* **2022**, *14*, 258. https://doi.org/10.3390/fi14090258

Academic Editor: Manuel José Cabral dos Santos Reis

Received: 28 July 2022
Accepted: 24 August 2022
Published: 30 August 2022

1. Introduction

Facial expressions are one of the most important ways in which humans can effectively communicate their emotional states and intentions [1,2]. The research on facial expression recognition has been enhanced, as more and more researchers focus their attention on the problem in the field of facial expression recognition. Automatic analysis techniques for facial expressions are being used in more and more fields, such as medical care, driver fatigue, robot interaction, and student classroom state analysis [3–7]. The changing needs have also given rise to many problems in different scenarios, and in the past period, the algorithms have been iterated continuously in FER-related problems [8–18], and eventually, these algorithms have achieved good results. Meanwhile, in order to meet the requirements of different experiments, some datasets have been generated, such as CK+ [19], FER2013 [20], FERPLUS [21], etc.

The construction of previous facial expression recognition systems is mainly divided into three steps: face detection, feature extraction, and expression recognition [22]. In dealing with facial expression recognition, the Dlib detector [23] is commonly used for face detection under various noise conditions. It can correct the images containing faces so that the images perform accurate face alignment. For face feature extraction, due to increasing computational power in recent years, researchers have designed many deep neural network algorithms to extract these geometric and appearance features [24,25]. ResNet [26] is frequently seen in recent methods used for feature extraction. The feature maps obtained from the feature extractor are fed into the classifier used to match expression classes in facial expression recognition. Meanwhile, the great increase in arithmetic power has largely accelerated the development of deep learning [27,28] and the research process of problems in the field of face recognition.

Traditional facial expression recognition methods [29,30] use manually extracted features such as the Histogram of oriented Gradients (HOG) [31], Local Binary Patterns (LBP) [32] and scale-invariant feature transformations (SIFT) [33] to deal with FER [34–36]. However, these methods usually have disadvantages, such as a long time spent, high cost, and poor portability. Compared with the traditional methods, deep learning methods can learn more deep features, such as rotation angle changes and changes caused by illumination. Additionally, due to the rapid development of the Internet in recent years, large-scale datasets can be obtained from challenging real-world scenarios in a relatively easier way [37,38]. Deep learning methods benefit from these advantages and are more popular among researchers. Information on some datasets commonly used for facial expression recognition is shown in Table 1.

Table 1. The image resolution of common datasets for facial expression recognition.

Datasets	Time	Images	Resolution
CK+	2010	1308	640 × 490
FER2013	2013	35,887	48 × 48
FERPLUS	2016	35,887	48 × 48

Despite the great success of facial expression recognition, there are still some challenges, such as the noise generated by intra-class differences unrelated to expressions and the problem of overfitting due to insufficient training data caused by too small a dataset. Further, first, there are large intersubject differences in faces in the dataset due to age, gender, and race [22]. Intra-class differences become larger because of these uncertainties. The intra-class variance unrelated to expression recognition is shown in Figure 1. Second, experiments on deep neural networks require a large amount of training data. Insufficient training data can easily lead to overfitting of the model. Finally, directly using the global face as the input of the neural network will lose local detailed feature information [39]. Ekman [40] studied facial parts to find out the most important regions for human perception by masking the key parts, and the regions around the eyes and mouth are highly correlated with FER. From Kotsia et al. [41], based on Gabor features and human observer synthesis analysis of obscured FER, the masked mouth affects the accuracy of FER more than the masked eyes. Inspired by this, we propose a Dual Path Feature Fusion and Stacked Attention Network (DPSAN) to address the challenges of noise generated by intra-class differences unrelated to expression recognition in FER and the challenges of overfitting due to insufficient training data. DPSAN consists of a dual path feature fusion module and a stacked attention module. Based on studies related to the symmetry of facial expressions [11,16], we segment four regions from the aligned face image, left eye, right eye, left half of the mouth, and right half of the mouth. The four key regions are used to reduce the effects of irrelevant intra-class differences, such as facial contour and hairstyle. Additionally, the effect of noise from occlusions, such as the partial occlusion of the left or right eye, is reduced based on the symmetry of facial expression features in local regions. Meanwhile, existing ensemble learning studies show [42] that the combination of multiple networks outperforms a single network. Therefore, we designed a Dual path feature fusion module. The dual path feature fusion module of DPSAN divides the face image feature extraction into global and local face features for extraction. The dual path feature fusion module fuses the global and local features extracted using the two updated ResNet18 [43]. By combining the global and local features, this module is used to overcome the noise caused by intra-class differences unrelated to expression recognition, and to increase the data to alleviate the problem of insufficient data. Additionally, multi-scale expression details are introduced in this module to improve the generalization ability of the model by adjusting the size of the local region to match the size of the global face image. Second, we design a stacked attention module for compressing the feature maps, reducing redundant data and supervising DPSAN, and also refine the global features via the local features of faces, enabling the model to learn more meaningful weights from both local- and global-based feature maps. In summary, our contributions are summarized as follows:

- We propose a DPSAN model to address the challenge of noise generated by intra-class differences in FER that are not related to expression recognition, and the challenge of overfitting due to insufficient training data.
- The dual path feature fusion module is able to segment the critical regions in the face and combine the local feature information with the global feature information. This can reduce the noise impact from intra-class differences in FER that are not related to facial expressions. Additionally, it can alleviate the overfitting problem caused by insufficient data. The stacked attention module is used to compress the feature maps and learn more meaningful weights from the local- and global-based feature maps.
- We perform extensive validation of our model on the dataset. The experimental results show that DPSAN outperforms previous state-of-the-art methods on the commonly used CK+ and FERPLUS datasets (93.26% on CK+, 87.63% on FERPLUS).

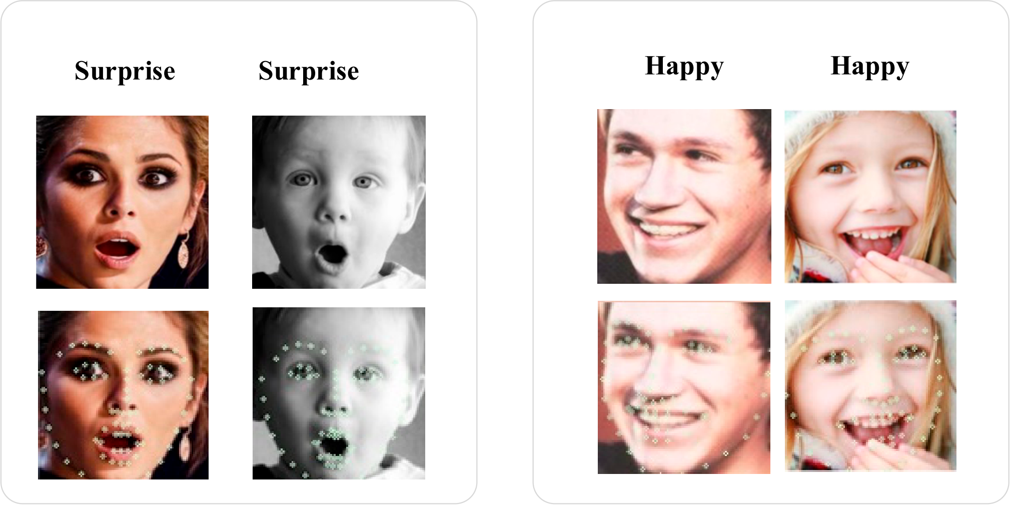


Figure 1. Intra-class differences are unrelated to expressions (the facial landmarks are detected by [44]). Images can have huge differences within the same expression class. For example, age and gender are different in different samples. As shown in the figure, the left and right images both represent the expressions of surprise and happiness, respectively, but they are visually very different. The above images are from AffectNet [7].

The rest of this paper is organized as follows. In Section 2, we review the literature related to the development of techniques related to facial expression recognition. In Section 3, we describe the research methodology in detail. In Section 4, we describe the experimental setup and provide a discussion of the experimental results. Finally, in Section 5, we conclude the paper and give an outlook on future work.

2. Related Work

Deep learning and Attention in FER: With the rapid development of computer vision in recent years, deep learning solutions are increasingly used to handle challenging tasks in FER compared to traditional manual feature extraction solutions. Meanwhile, traditional CNNs pay insufficient attention to important channel features and important regions of images, which restrict the accuracy of facial expression recognition. However, with the proposed attention, the model will give more attention and weight to important channels and regions, which is a good improvement to the deficiency of the original traditional model. Many existing models of deep learning and attention have achieved good performance. Li et al. [11] proposed a convolutional neural network with attention (ACNN) to automatically perceive blocked face regions and focus primary attention on unblocked expression regions for judgment. Wang et al. [16] proposed a regional attention network (RAN) to recognize the occlusion of facial regions. Wang et al. [15] proposed a self-cure network (SCN) to suppress bad data such as obscured face images in many large datasets, which prevents the network from overfitting incorrectly labeled samples.

Facial Landmark in FER: The purpose of facial landmark detection is to identify the location of distinguishable key points on a human face. Currently, with the combination of facial landmark detection and deep learning techniques, such as face recognition [15,20], facial expression recognition [13,31], and face tracking [17], great progress has been made in the application of facial landmark detection, which has also facilitated the research of many accurate facial landmark detectors. Many accurate and widely used facial landmark detectors have been proposed, such as [4,11,14,45]. Using these excellent detectors, researchers are able to better utilize facial landmarks as informative features in solving FER tasks.

Ensemble learning in FER: Previous studies on ensemble learning have amply demonstrated that the combination of multiple networks outperforms individual networks [42]. When implementing network combinations, two key factors should be considered: (1) to ensure complementarity among multiple networks, we can increase the diversity of networks; and (2) choosing the right combination strategy can provide more effective assistance in aggregating committee networks [22]. For the second point, each member of the committee network can be combined at the feature level and at the decision level, respectively. For the aggregation of features, concatenating features obtained in different network paths is the most commonly used strategy [12,46]. For the decision level, there are several most frequently used rules, such as simple averaging, majority voting, and weighted averaging. Yovel et al. [39] showed that people could effectively access the information displayed by facial expressions through two paths: local regions of the face and the whole face.

Compared with previous methods, our proposed method is very effective in reducing the interference of irrelevant intra-class differences using a unique four key region segmentation method. Additionally, the method has good robustness due to the fixed region cropping approach. Comparing previous methods on the same dataset, the proposed method in this paper effectively expands the data volume via cropping segmentation to obtain detailed information in local paths while better meeting the experimental requirements of the dataset. The dual path feature fusion module and stacked attention module designed in this paper can make better use of contextual information and also focus more attention on more important regions.

3. Methods

3.1. Overview

As shown in Figure 2, first, the face image is segmented using the segmentation method to segment four local image regions. The face key region segmentation method can extract more local details, and the size of each region image is resized to be equal to the size of the global input image via scaling. Then, the global image and the four local image regions are input to the updated ResNet18 for feature extraction in the global path and local path, respectively. The general structure of the updated ResNet18 is shown in the dashed rectangle at the lower left corner, and a more detailed structure is shown in Figure 3. After completing the feature extraction, the global image feature map is fused with the local image feature map. After that, the fused feature maps are fed into the stacked attention module (SA). The SA module can learn features from different regions and complete the weighting of different channels by re-estimating the importance of each feature. The image segmentation method is shown in Figure 4. The details of the structure of the stacked attention module are shown in Figure 5.

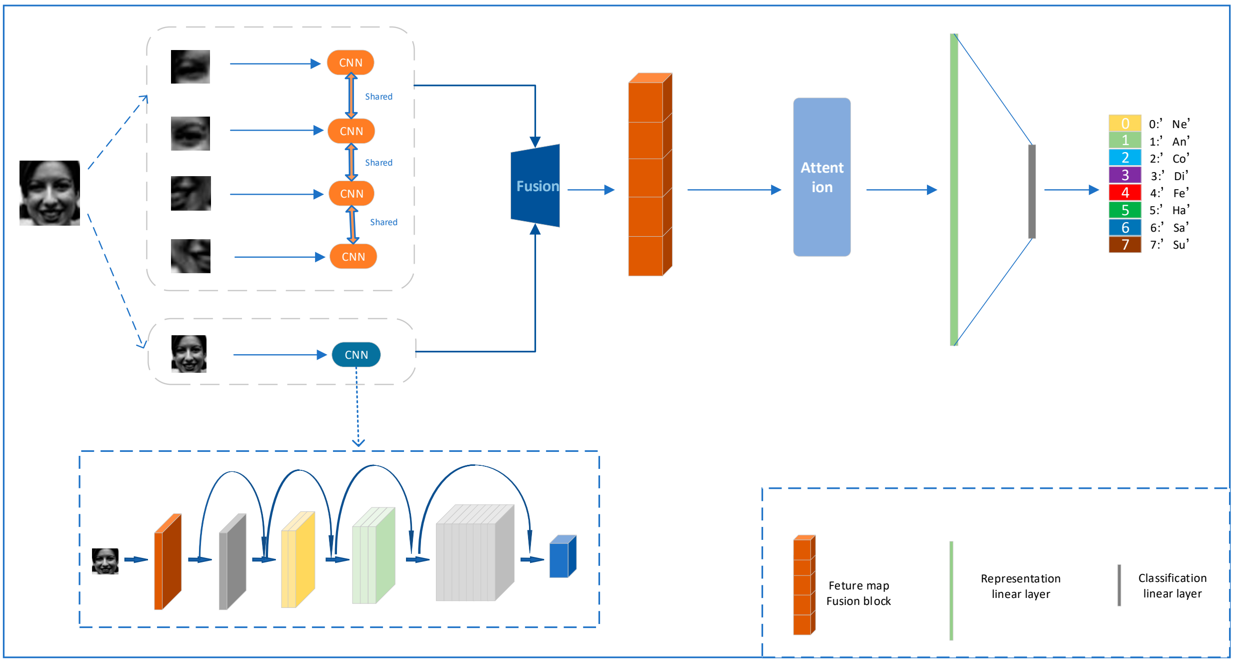


Figure 2. The framework of DPSAN consists of a dual path feature fusion module and a stacked attention module.

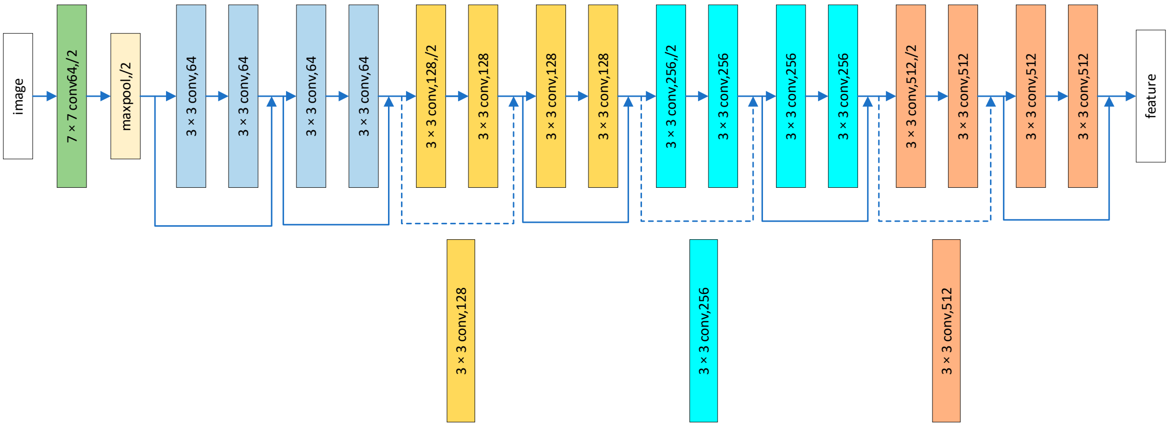


Figure 3. The updated ResNet18. We removed the avg-pool layer and the fully connected (FC) layer to facilitate the next step of feature fusion.

The model consists of two main parts. One is the dual path feature fusion module, which performs segmentation, feature extraction, and feature fusion on the input face image. The other is the SA module, which weights each channel of the feature map according to its contribution to the classification result. The detailed process of the first module is as follows. First, the model segments the local area of the base image, and compresses and scales the size of the local area image and the size of the base image; and then the scaled image is used as the input to the network. The dual path feature fusion network concatenates the features extracted from the four local regions with the features of the complete face image and outputs the fused feature map. The main role of the SA module is to calculate the importance of the feature maps of each layer and compress the irrelevant information. Finally, the classification of expressions is output from the linear connectivity layer. Figure 2 illustrates the DPSAN framework.

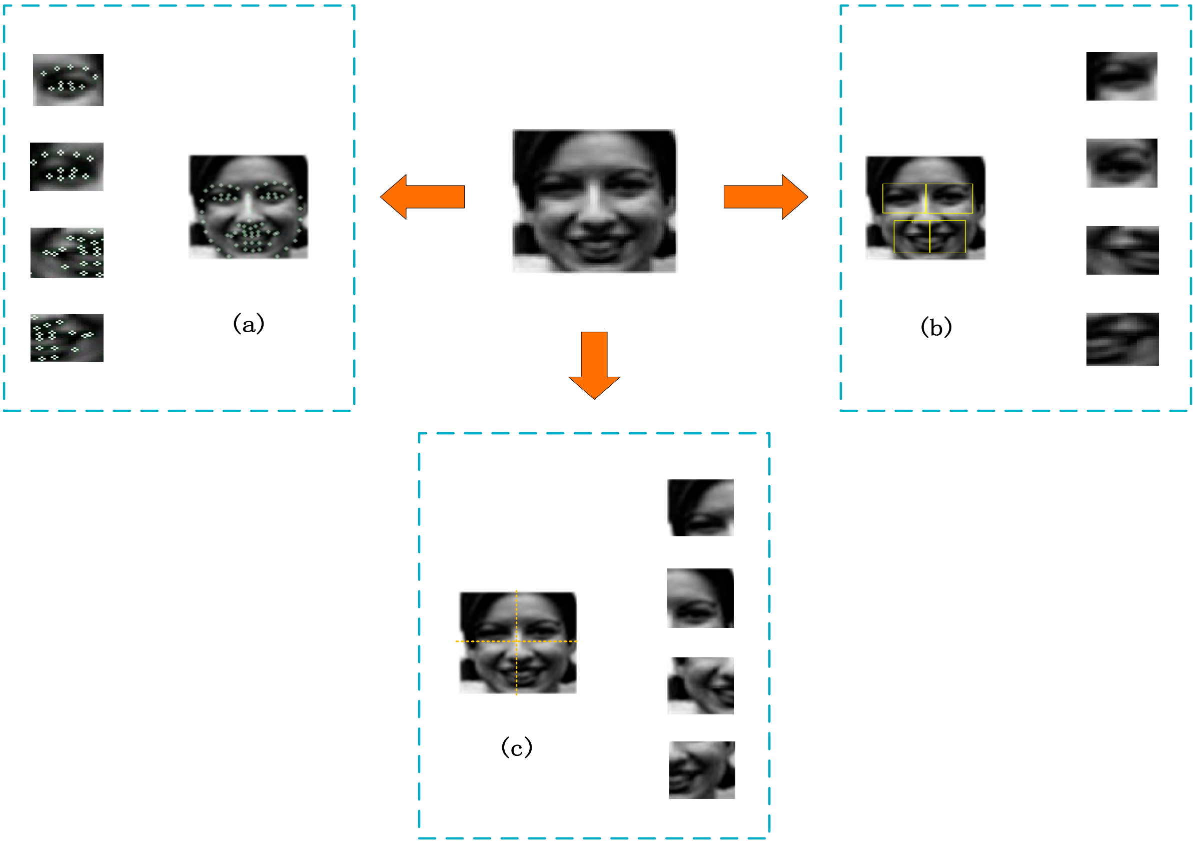


Figure 4. Example of facial key region segmentation method. (**a**) The 42-point cropping method based on face; (**b**) Crop method based on four regions of the face; (**c**) Four-equivalent cropping method based on face image.

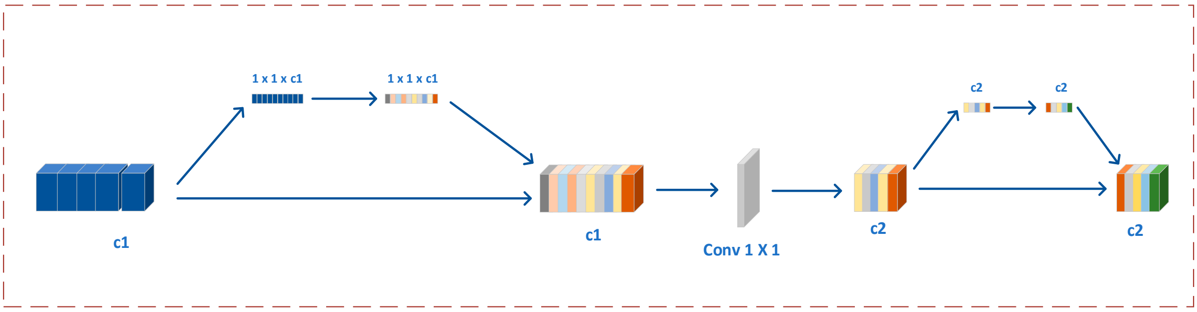


Figure 5. SA consists of two channel attention networks and a 1 × 1 convolutional neural network conv 1 × 1. DPSAN accomplishes the weighting of five different regions through the channel attention mechanism.

DPSAN is used with one global image and four local image regions as inputs. These five parts are input to two ResNet18, and the features are extracted and represented as five sets of feature maps. Then, DPSAN merges the global feature maps and the four sets of local feature maps into one set of feature maps. The advantage of this processing is that the fused feature maps have both the global information of the facial image and local information that can be easily ignored. The SA module calculates the weights for each channel of the feature map. The weighted features contain all the key information that we are interested in that is beneficial to FER. The classification layer after the stacked attention module aims to classify the face images into their corresponding facial expression classes. The softmax loss is used to optimize the DPSAN model.

3.2. Face Image Segmentation and Dual Path Feature Fusion Module

The face image segmentation and dual path feature extraction module firstly segment the input base image, crop it to obtain four local regions, and then extract the global features and local features from the adjusted global image and the four local region images through the dual path, respectively. The advantage of dual path feature extraction is that the global information is obtained along with the local detail information, thus complementing the information ignored by the global information. Then, the global features and the local part of the features are fused, which can make the feature representation more adequate and complete.

We extract feature information from the base face image and the local image regions, using the updated ResNet18 network. The detailed information on the updated ResNet18 is given in Figure 3. In the updated ResNet18, the avg-pool layer and the FC layer in the structure are removed in this paper to facilitate the computation and fusion of features later. For the segmentation methods, specifically, we propose three methods for face region segmentation for comparison, and then we select a better method from them to segment the global image and obtain the local block.

Comparison and selection of face key region segmentation schemes. With the study conducted by Ekman [40] on facial parts, this paper extracts the features of specific regions related to facial expressions, which contribute to FER. First, we use Dlib, which is popular among researchers, to detect facial regions. Then, the eyes, nose, and left half of the mouth, along with the right half of the mouth, are selected from the facial regions as patches that are highly relevant to FER. Figure 4 shows the different methods for the selection of facial regions. The three methods are as follows:

(1) **Forty-two-point cropping method based on face:** (a) After we read and studied many papers, we used the method from [47] to detect 68 landmark points in the facial image after selection. Then, we selected 42 points that could completely cover the eyes, nose, left half of the mouth, and right half of the mouth. (b) Eleven point pairs are selected around the eyes, and the maximum and minimum coordinates of each eye are calculated as rectangular boundaries. We select a total of 20 points around the mouth, with 12 points each for the left and right half of the mouth, of which four points are duplicated. The two local regions are divided by the four points located in the middle for the left and right half of the mouth, and then the maximum and minimum coordinates of the left and right parts of the mouth are calculated as the rectangular boundaries, respectively. The image of the segmented region is shown in Figure 4.

(2) **Crop method based on four regions of the face:** The local area of the face image is relatively regular. Four face images can be dropped at a key fixed location and the coordinates of the rectangular frame can be directly calculated to obtain four local blocks.

(3) **Four-equivalent cropping method based on face image:** The fixed size of the input face image is 96×96. The midpoint coordinates of the length and width of the face image are calculated and divided directly into four local blocks, with a size of 48×48.

After region segmentation, the next step is to extract features from the global image and local image regions. Since the features of the global image tend to ignore the attention to local detail features, considering the different attentions to the highly correlated local region image and the global structure image, the DPSAN in this paper introduces two paths: the local path network (LPN) and the global path network (GPN). First, the local region face images and global face images are resized to 96×96. As shown in Figure 2, the LPN consists of local blocks with the resized four sets of local face images as the input. The GPN consists of only the resized global face images as the global input patches. Then, the five sets of feature maps are combined into one set. The LPN is designed to focus on highly relevant detailed information that is ignored by the global patches. The Dlib face detector is commonly used to locate faces in images. For images containing only faces detected using the Dlib detector, a face region segmentation module is required to obtain four regions. In this paper, details are given in Figure 2.

As shown in Figure 2, four local regions of the eyes, nose, left half of the mouth, and right half of the mouth are selected from the original image and resized to 96 × 96. Using only global images or local images will ignore some associated information in the face image. Feature fusion of the global image and local image regions leads to a better performance. DPSAN considers both the global face and local face blocks. First, considering the difference in information between the global and local images, this paper encodes the feature maps of global face and local face blocks in two ResNet18 separately so that DPSAN obtains facial expression information from the local features and global features, respectively, and refines the contextual information that may be overlooked. Second, ensemble learning studies show that multiple network combinations outperform individual network structures, and acquiring features from different scales globally and locally can also help the network to learn more diverse information related to expressions. The images of the global face regions and local face blocks are input into two ResNet18, and the final feature map of 5 × 512 × 21 × 21 is obtained after transformation.

3.3. Stacked Attention Module

The stacked attention (SA) module computationally assigns different weights to each channel feature map with the purpose of learning weights that are more meaningful for expression recognition from the fused local and global feature maps. This module compresses the fused feature maps and filters out redundant information. In this paper, we embed SA after the dual path feature fusion module based on two ResNet18 networks to automatically reweight the whole feature map. Our proposed SA consists of two channel attentions and a 1 × 1 Conv. The detailed structure of the SA is shown in Figure 5.

The role of channel attention is to assign weights to different channels of the feature map. In the SA module, a fused feature map F_f is first computed with F_g and F_{l_i} as the inputs. Then, a F'_f is sequentially obtained after a channel attention map M_{c_1} [48], with F_f being computed with the feature map F_f as input. After a 1 × 1 convolutional neural network calculation, another channel attention map, M_{c_2}, is used to calculate F''_f, as shown in Figure 5. The functions of the SA module are described as follows.

$$F_f = concat(F_g, F_{l_1}, F_{l_2}, F_{l_3}, F_{l_4}) \tag{1}$$

$$F_f{}' = M_{c_1}(F_f) \otimes F_f \tag{2}$$

$$F_f{}'' = M_{c_2}(conv_{1\times 1}(F_f{}')) \otimes F_f{}' \tag{3}$$

where F_g is the feature map extracted from the global path in the dual path fusion module based on the ResNet18 network, F_{l_i} is the feature map extracted from the local path, M_{c_i} is the channel attention network, and ⊗ indicates multiplication. F'_f is the output after the first channel attention calculation, and F''_f is defined as the final output.

For the facial expression feature maps, the weights learned for each channel dimension are different. In the process of facial expression recognition, features are extracted from each local region and global image. By weighting each channel feature, we can assign different weights to the feature maps from the four different regional images and the global image, thus weighting the fused overall feature map. The detailed structure of SA is shown in Figure 5.

There are two steps in SA. The first step is to use the attention network to learn weights to represent the importance of the feature maps, and then merge the five groups of feature maps into one group. After our validation, the feature maps obtained with only one channel attention are too large for direct classification and do not facilitate back-propagation. The features extracted from the global and local regions are too large, and these data need to be compressed and further learned. Therefore, in the next step, a 1 × 1 convolution and another attention network are used to reduce the feature map dimensions and reweight their importance to facilitate the next step of encoding them into feature vectors corresponding to facial expressions.

3.4. Loss Function

In the experiments of this paper, to compare the difference between the predicted labels and the real labels, and to minimize the differences between them, we use the cross-entropy loss function. This function is as follows:

$$L = -\sum_{i}^{N}\sum_{c=1}^{8} y_c^i \log(p_c^i) \tag{4}$$

where y_c^i denotes the indicator variable, and the value of y_c^i is 1 if the true label of sample i is equal to c; otherwise, y_c^i is 0. p_c^i is the predicted probability that the observed sample i belongs to category c.

3.5. Summary

In order to improve the accuracy of facial expression recognition, this paper proposes improvements to the problems of insufficient attention to local detail features and irrelevant noise interference that exist in the previous FER models. First, we propose a dual path approach to segment the global face image and to extract local detail features to complement the global features via local paths, which improves the problem of insufficient attention to local features in the previous FER models and ignores the interference of irrelevant information for expression recognition. Second, we also propose the SA module to calculate the importance of each channel feature map in the fused feature map and also compress redundant data to improve the computational efficiency of DPSAN.

4. Experiment

In our experiments, we will use two public datasets, the CK+ dataset and the FERPLUS dataset. Then, the effectiveness and robustness of DPSAN will be demonstrated through experiments on these two datasets. Then, DPSAN is compared with some of the latest methods available on the CK+ and FERPLUS public datasets, respectively, to demonstrate the advancedness of DPSAN. We also designed ablation experiments to demonstrate the effectiveness of each module in DPSAN by splitting the dual path feature module and the stacked attention module in separate comparison experiments.

4.1. Datasets

Two public datasets, CK+ [19] and FERPLUS [21], will be used in the experiments, and the experimental data on these two datasets will be used to analyze and evaluate our proposed method. These two datasets contain many different numbers of face images. In Figure 6, we show some examples of differently labeled images from these two datasets. The public datasets of facial expressions used in the paper are used in accordance with the official requirements of the datasets. In accordance with these official statements, we have cited the corresponding literature in the paper.

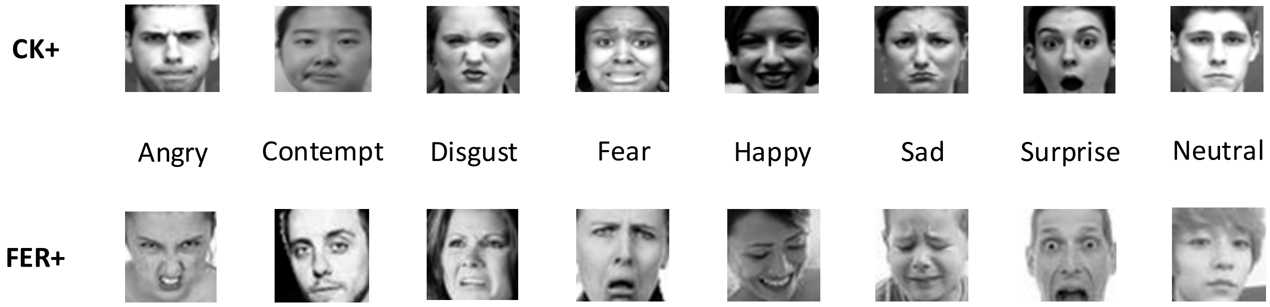


Figure 6. Some examples of the dataset used in the experiment. The first row of images is from CK+, and the second row is from FERPLUS, which has eight emotions (anger, contempt, disgust, fear, happiness, sadness, surprise, and neutrality), with the new emotion label of contempt added to FERPLUS.

The extended Cohn-Kanade (CK+) [19] dataset is the most widely used dataset in studies of facial expression recognition. One hundred and twenty-three subjects were included in the CK+ dataset, in which 593 video sequences were recorded. Based on the Facial Action Coding System (FACS), the peak expressions in the 327 video sequences of 118 subjects were labeled with seven basic expression labels, which were happy, angry, sad, disgusted, fearful, surprised, and contempt. Additionally, the first frame image in the video sequence was usually used as the neutral expression, which is the eighth basic expression label.

The FER2013 [20] dataset was first introduced in the 2013 ICML challenge. This dataset is a dataset automatically collected from the web using the Google search engine without any constraints. The dataset contains a large number of facial expression images, which are adjusted and cropped to a resized image size of 48 $*$ 48. The FER2013 dataset consists of 28,709 training images, 3589 test images, and 3589 validation images. The dataset FERPLUS [21] used in our experiments relabeled the original images in the FER2013 dataset and added the emotion label of light contempt. Additionally, each image had 10 taggers, and the labeled data allowed us to better estimate the emotional probability distribution of each face image. The eight emotion labels in the FERPLUS dataset were happiness, disgust, fear, sadness, anger, surprise, neutrality, and contempt, where the contempt tag is newly added.

4.2. Implementation Details

Pytorch [49] and TensorFlow [50] are both excellent open source frameworks in deep learning. Our experiments use the Pytorch framework to implement DPSAN. Previous experiments have demonstrated that ResNet18 [43] can show excellent performance in deep learning, so two ResNet18 networks are used for feature extraction in the dual path feature fusion block of DPSAN. The batch size is set to 8, the maximum number of iterations is 1200, the learning rate is 0.0001, the Stochastic Gradient Descent (SGD) is used to optimize the network parameters, and finally, all other parameters are set to default values. Regarding the hardware parameters in the experiments, the DPSAN experiments were performed on an RTX 2080Ti GPU with 11 GB of RAM.

In order to improve the computational efficiency in the experiment, we designed some operations that can increase the computational speed. For the CK+ dataset, since the original image size in this dataset is 640 × 490, the information in the image has some information that is irrelevant for FER. For the global face image, we use the Dlib face detector to crop out the face image and resize the cropped image to 96 × 96, which can exclude the intra-class differences irrelevant to facial expressions and reduce the computational effort at the same time. For the local region blocks, we also cropped the left eye, right eye, left half of the mouth, and right half of the mouth from the original image and resized these regions of interest to 96 × 96. After the above operations, we evaluated the accuracy and speed of the three cropping methods via 10-fold cross-validation on the CK+ dataset. These are the crop method based on 42 points of the face, the crop method based on four fixed regions of the face, and the crop method based on four equal parts of the face image. The local area images cropped using the three methods are adjusted to the same size in the experiments. The parameter settings in the experiments are kept the same as the default settings. We show the results of the three cropping methods in Figure 7, and the data show that the cropping method based on the fixed four regions of the face can obtain the regional information related to facial expression recognition faster and more effectively. First, compared with the four-equivalent cropping method based on face image, the cropping method based on four fixed regions of the face can obtain the FER-related region information more effectively; meanwhile, compared with the cropping method based on 42 points of the face, the cropping method based on four fixed regions of the face can obtain the FER-related region information faster because the cropping is performed almost instantaneously. More importantly, the face fixed four region-based cropping method 10-fold cross-validation is also higher than the previous two, in terms of accuracy. Therefore, the experimental cropping method used in the dual path feature fusion module is based on the face fixed four regions cropping method.

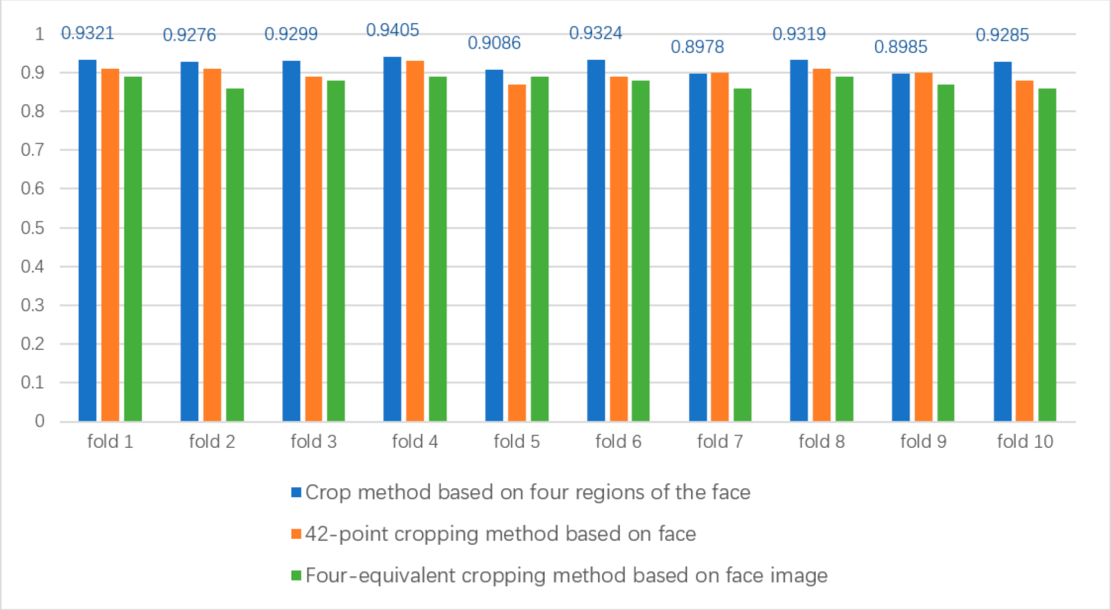


Figure 7. Accuracy (%) of 10-fold on the CK+ validation set using three facial region segmentation methods.

Previous ensemble learning studies have shown that the combination of multiple networks can learn more effective information from images due to a single network. Good results have been obtained for integration learning in many areas of research, such as [51–53]. One of the feature combinations for emotion recognition in [46] is fc5 VGG13 + fc7 VGG16 + pool ResNet, which has the disadvantage of the network and parameter size being too large and not easy to train. Our experiments select two ResNet18 for feature extraction, where four local regions share the same ResNet18, thus achieving improved computational efficiency while enabling local region images to provide more effective feature information for facial expression recognition. Meanwhile, due to the excessive number of features obtained from the dual path feature fusion module, the feature vector created by these features is relatively inefficient and difficult to train after only one channel attention network. So, we add a 1 × 1 convolution here, which reduces the number of parameters. The stacked attention module allows for reweighting the importance of each channel in the feature map while compressing the feature map and reducing the number of parameters through a 1 × 1 convolution and another channel attention network.

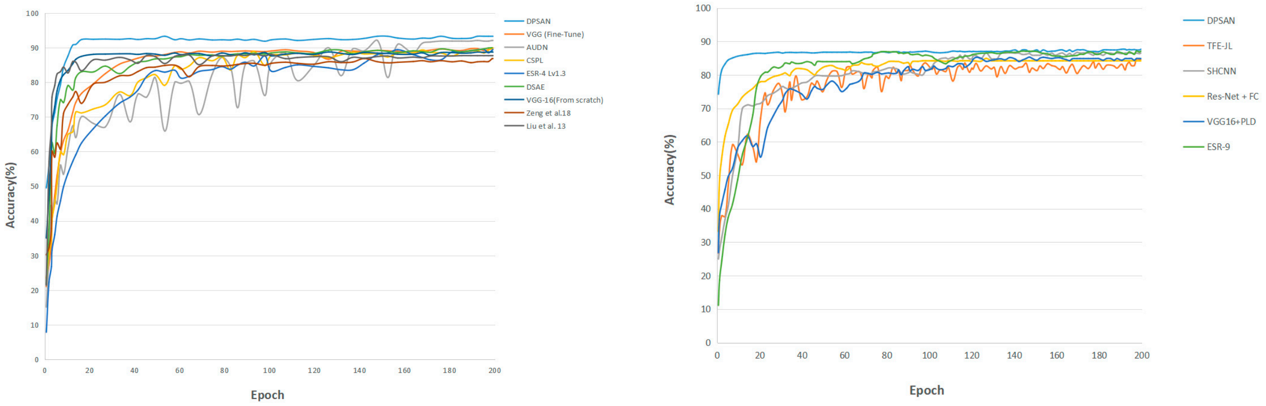
4.3. Comparison of DPSAN on the CK+ dataset

Combining the characteristics of the CK+ dataset, we first extracted the first frame from each sequence on the CK+ dataset as a natural neutral expression while labeling the last three frames of the sequence with the peak expression as one of the seven basic emotions. Then, the global face image is cropped using the dlib face cropper, and four local image regions of the eyes, nose, left half of the mouth, and right half of the mouth are segmented using the face key region segmentation method. The cropped images are then resized to reset the image size to 96 × 96 pixels. Since the CK+ dataset does not provide a specified training and validation set, we choose the 10-fold cross-validation scheme for training and validation. Finally, the average cross-validation recognition accuracy of 10 times is taken as the result. In this paper, we show the accuracy of the validation set of the CK+ dataset in Figure 7, and the selected method is labeled with the data in Figure 7.

In Table 2, this paper compares DPSAN with a variety of state-of-the-art methods on CK+. Among them, the accuracy of AUDN and VGG (Fine-Tune) is 92.05% and 89.90%, respectively, our proposed DPSAN reaches 93.26%, and the data in the table show that DPSAN outperforms these state-of-the-art methods. The results of DPSAN compared to other methods on the CK+ dataset are shown in (a) of Figure 8.

Table 2. Comparison of DPSAN with other models on the CK+ dataset.

Method	Average Acc (%)
ESR-4 Lv1.3 [14]	89.4
AUDN [54]	92.05
DSAE [55]	89.84
VGG-16 (Fine-Tune) [56]	89.90
CSPL [30]	89.89
VGG-16 (From scratch) [56]	88.70
Zeng et al. [55]	86.82
Liu et al. [54]	87.67
DPSAN (Ours)	**93.26**



(**a**) (**b**)

Figure 8. Results of DPSAN with other methods are shown. (**a**) Results on the CK+ dataset; (**b**) Results on the FERPLUS dataset.

4.4. Comparison of DPSAN on the FERPLUS Dataset

FERPLUS is an extension of the original FER dataset. In order to better test the recognition effect of DPSAN in the real environment in the field, this paper conducts training validation on the FERPLUS dataset. Since the image size in the FERPLUS dataset is 48 × 48, the image size is resized to 96 × 96 pixels in this paper. The DPSAN model is faster than CK+ on the FERPLUS dataset, so the initial learning rate is set to 0.1, and it is updated by multiplying it by 0.75 after every 10 cycles. Other settings remain the same as CK+.

We show the results of the DPSAN comparison experiments on the FERPLUS dataset in Table 3. In this paper, we compare DPSAN with various state-of-the-art methods successively. Additionally, we show the confusion matrix of DPSAN sentiment prediction on the FERPLUS dataset in Figure 9. By comparing DPSAN with multiple advanced methods on FERPLUS, where the accuracy of ESR-9 and SHCNN is 87.15% [14] and 86.5% [57], respectively, the accuracy of the DPSAN proposed in this paper is 87.63%, which shows that the effectiveness of our DPSAN on FERPLUS is up to the current advanced level. The results of DPSAN with other methods on the FERPLUS dataset are shown in (b) of Figure 8.

Table 3. Comparison of DPSAN with other models on the FERPLUS dataset.

Method	Average Acc (%)
TFE-JL [58]	84.3
SHCNN [57]	86.5
Res-Net + FC [59]	83.4
VGG16 + PLD [21]	84.99
ESR-9 [14]	87.15
DPSAN (Ours)	**87.63**

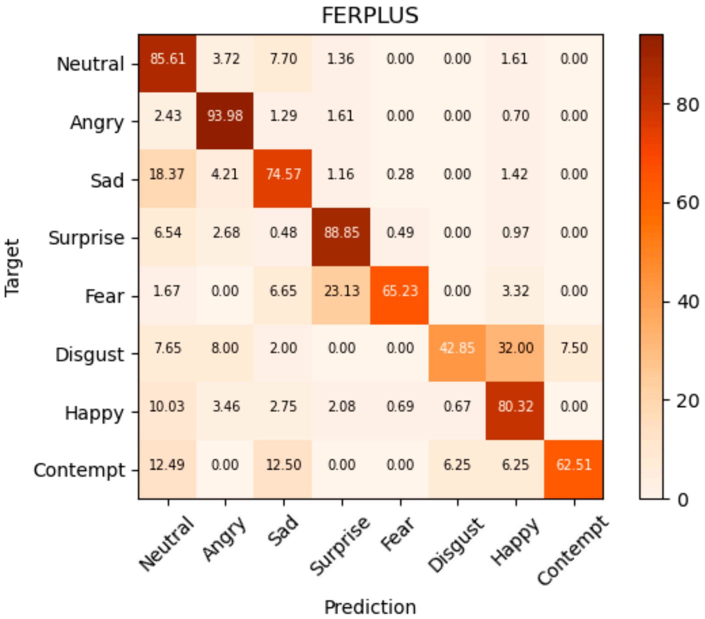


Figure 9. Confusion matrix for DPSAN prediction based on FERPLUS dataset and sentiment label distribution.

4.5. Ablation Experiments

In order to fully investigate the effectiveness of each module of DPSAN, an ablation experiment is designed in this paper to verify the effects of two modules, the dual path feature fusion module and the SA module, on FER.

The baseline model global path network (GN) is designed in the experiment. This model only takes the global image as input and generates a feature map through a ResNet18 shown in Figure 3, and then classifies the expressions via a classifier. In order to demonstrate the advantages of the dual path feature fusion module and the SA module, two groups of comparative experiments are designed; the comparison between the dual path network (DPN) and GN, and the comparison between DPN and DPSAN.

First, to demonstrate that dual path networks can reduce model overfitting and have better accuracy stability over different data, a DPN model is experimentally designed. The DPN model has one more local path than the GN model, which takes the global image and local image blocks as input and is compared with GN for both accuracy and standard deviation. Second, to demonstrate the advantages of the SA module, DPN is compared with DPSAN, in which DPN lacks the stacked attention module. By comparing the two, the superiority of DPSAN in both accuracy and standard deviation can be demonstrated.

The histogram of the comparison between the two groups of experiments is shown in Figure 10. We annotated the experimental data of DPSAN. From Figure 10, it can be roughly seen that the accuracy of GN fluctuates a lot in 10-fold cross-validation, while DPN fluctuates less, indicating that it is less stable than DPN; and at the same time, the advantage of accuracy is too small. The performance of DPSAN clearly exceeds that of the baseline, i.e., GN. The experimental results are shown in Table 4.

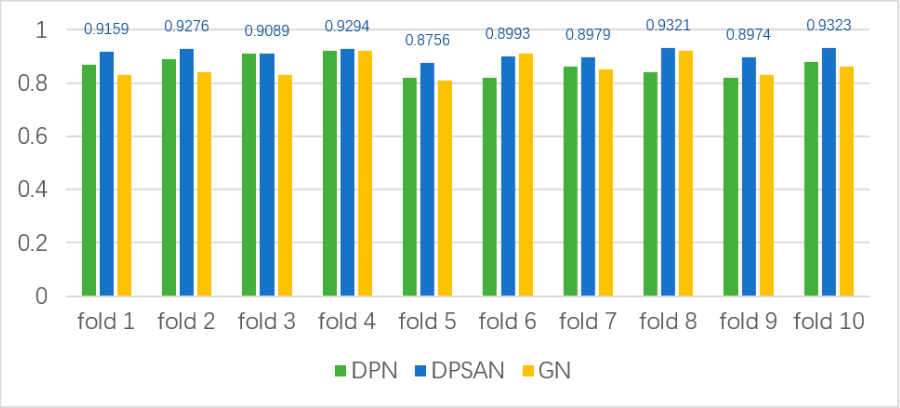


Figure 10. Performance of different component solutions on CK+.

Table 4. Evaluation of all components of DPSAN on the CK+ dataset.

Global Path	Local Path	SA	Average Acc (CK+)
√	×	×	89.74 (±4.7)%
√	√	×	89.23 (±3.6)%
√	√	√	93.26 (±3.5)%

(1) DPN VS. GN: DPN and GN were compared in the experiments to demonstrate the benefits of adding a local path. The combined results in Figure 10 and Table 4 can now show that GN performs slightly better than DPN in terms of accuracy in certain folds. However, the standard deviation calculated by DPN is much smaller than the results of GN, which is more stable than GN, and which also indicates that DPN can introduce multi-scale expression details by adding local paths in facial expression recognition through region segmentation, and that it can complement the detailed attributes ignored by the global image features and effectively avoid overfitting. The reduction in the standard deviation shows that combining local details and global feature mapping can well suppress intra-class differences.

Additionally, considering the different contributions of local regions and global images to the FER problem, these features need to be re-weighted. Here, we design the stacked attention module.

(2) DPSAN VS. DPN: Experiments were conducted to compare DPSAN and DPN to validate the benefits of overlaying the attention module. As shown in Table 4, the accuracy of DPSAN improved by 4.03% on the CK+ dataset. As shown in Figure 10, DPSAN outperforms DPN and GN in almost all folds. The comparative experimental data in Table 4 show that the SA module in DPSAN can effectively reassign the weights of each feature map based on its contribution to FER.

Through two sets of ablation experiments, the following conclusions can be drawn. First, it can be demonstrated that the dual path feature fusion module in the DPSAN model is effective and can effectively suppress the intra-class variance, but it leads to a slight decrease in recognition accuracy. Second, DPSAN achieves good results after the addition of the stacked attention module. The model achieves a large improvement on the CK+ dataset, and the accuracy improves from 89.76% to 93.26%, which indicates that the stacked attention module is an important contributing module of DPSAN, and also proves the effectiveness and advancement of the proposed stacked attention module.

5. Conclusions

In this paper, we propose a network model DPSAN for facial expression recognition. A dual path feature fusion module and a stacked attention module are designed in this paper. The model utilizes more detailed information that is ignored by the global image, reduces the effect of intra-class differences that are not related to expression recognition, and also expands the size of the experimental dataset.

The dual path feature fusion module designed in this paper by combining the global and local features can be used to reduce the impact of noise caused by intra-class differences unrelated to expression recognition, and to increase the data to alleviate the problem of insufficient data. In order for DPSAN to learn key features more effectively, compress features, and reduce the number of parameters, a stacked attention module is designed in the model and it achieves better results in terms of the accuracy of FER. Finally, the analysis of ablation experiments shows that both the dual path network and the stacked attention module are important modules of DPSAN. Experiments under two dataset evaluation protocols show that DPSAN achieves the performance of other state-of-the-art methods. At the same time, due to the presence of some images in the CK+ and FERPLUS datasets that do not have obvious expression features or images with relatively obvious intra-class differences, these factors can make it difficult for the image algorithm to obtain a very good performance gain at once. Therefore, the accuracy improvement obtained by DPSAN is meaningful when compared with the accuracy improvement obtained using other methods.

There are also shortcomings in this paper. In the experiments, our model only considers expression recognition on a pure dataset of face expressions without occlusions. However, in real scenarios, face images may be obscured by glasses, masks, and many other occlusions. Especially in the current global epidemic, wearing masks has become the norm in people's lives, and these occlusions can affect the accuracy of expression recognition. So, how to improve the accuracy of facial expression recognition under the condition of partial occlusions, such as mask occlusion and glasses occlusion, will be one of the major challenges to overcome in our research in the future.

In our future work, we will further investigate the following perspectives in order to reduce the impact of occlusion on expression recognition. First, we will consider the use of spatial attention for obtaining the feature information of unobscured regions and more important regions in face images from a global perspective. Second, considering the advantage of local paths to capture more details in this paper, we will consider segmenting the occluded face image into several local regions for feature extraction. Finally, we will consider using both channel attention and spatial attention to give higher weights to the features of important regions spatially and to the features of important channels spatially, so as to reduce the effect of occlusion on expression recognition.

Author Contributions: Supervision, H.X. and X.M.; Conceptualization, H.X. and X.M.; Formal analysis, H.X. and M.B.; Methodology, H.Z.; Writing—original draft, H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All datasets used in this paper are publicly available, AffectNet at http://mohammadmahoor.com/affectnet/ (accessed on 10 October 2021), CK+ at http://www.jeffcohn.net/Resources/ (accessed on 5 November 2021), FERPLUS at https://github.com/Microsoft/FERPlus (accessed on 19 November 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tian, Y.I.; Kanade, T.; Cohn, J.F. Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 97–115. [CrossRef] [PubMed]
2. Darwin, C.; Prodger, P. *The Expression of the Emotions in Man and Animals*; Oxford University Press: Oxford, UK, 1998.
3. Dhall, A.; Kaur, A.; Goecke, R.; Gedeon, T. Emotiw 2018: Audio-Video, student engagement and group-level affect prediction. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, Boulder, CO, USA, 16–20 October 2018; pp. 653–656.
4. Fabian Benitez-Quiroz, C.; Srinivasan, R.; Martinez, A.M. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5562–5570.

5. Dominguez-Catena, I.; Paternain, D.; Galar, M. Assessing Demographic Bias Transfer from Dataset to Model: A Case Study in Facial Expression Recognition. *arXiv* **2022**, arXiv:2205.10049.
6. Li, S.; Deng, W.; Du, J. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2584–2593. [CrossRef]
7. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Trans. Affect. Comput.* **2017**, *10*, 18–31. [CrossRef]
8. Cai, J.; Meng, Z.; Khan, A.S.; Li, Z.; O'Reilly, J.; Tong, Y. Island loss for learning discriminative features in facial expression recognition. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 302–309.
9. Hou, C.; Ai, J.; Lin, Y.; Guan, C.; Li, J.; Zhu, W. Evaluation of Online Teaching Quality Based on Facial Expression Recognition. *Future Internet* **2022**, *14*, 177. [CrossRef]
10. Sangermán Jiménez, M.A.; Ponce, P.; Vázquez-Cano, E. YouTube Videos in the Virtual Flipped Classroom Model using Brain Signals and Facial Expressions. *Future Internet* **2021**, *13*, 224. [CrossRef]
11. Li, Y.; Zeng, J.; Shan, S.; Chen, X. Occlusion Aware Facial Expression Recognition Using CNN with Attention Mechanism. *IEEE Trans. Image Process.* **2018**, *28*, 2439–2450. [CrossRef]
12. Liu, K.; Zhang, M.; Pan, Z. Facial expression recognition with CNN ensemble. In Proceedings of the 2016 International Conference on Cyberworlds (CW), Chongqing, China, 28–30 September 2016; pp. 163–166.
13. Roy, S.; Etemad, A. Analysis of Semi-Supervised Methods for Facial Expression Recognition. *arXiv* **2022**, arXiv:2208.00544.
14. Siqueira, H.; Magg, S.; Wermter, S. Efficient Facial Feature Learning with Wide Ensemble-Based Convolutional Neural Networks. In Proceedings of the AAAI Conference on Artificial Intelligence, Palo Alto, CA, USA, 3 April 2020; Volume 34, pp. 5800–5809. [CrossRef]
15. Wang, K.; Peng, X.; Yang, J.; Lu, S.; Qiao, Y. Suppressing uncertainties for large-scale facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6897–6906.
16. Wang, K.; Peng, X.; Yang, J.; Meng, D.; Qiao, Y. Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition. *IEEE Trans. Image Process.* **2020**, *29*, 4057–4069. [CrossRef]
17. Yu, Z.; Zhang, C. Image Based Static Facial Expression Recognition with Multiple Deep Network Learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; ACM: New York, NY, USA, 2015; pp. 435–442.
18. Gloor, P.A.; Fronzetti Colladon, A.; Altuntas, E.; Cetinkaya, C.; Kaiser, M.F.; Ripperger, L.; Schaefer, T. Your Face Mirrors Your Deepest Beliefs—Predicting Personality and Morals through Facial Emotion Recognition. *Future Internet* **2022**, *14*, 5. [CrossRef]
19. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.
20. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Bengio, Y.; et al. Challenges in representation learning: A report on three machine learning contests. In Proceedings of the International Conference on Neural Information Processing, Bangkok, Thailand, 18–22 November 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 117–124.
21. Barsoum, E.; Zhang, C.; Ferrer, C.C.; Zhang, Z. Training deep networks for facial expression recognition with crowd-sourced label distribution. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; pp. 279–283. [CrossRef]
22. Li, S.; Deng, W. Deep Facial Expression Recognition: A Survey. *IEEE Trans. Affect. Comput.* **2020**. [CrossRef]
23. Amos, B.; Ludwiczuk, B.; Satyanarayanan, M. Openface: A general-purpose face recognition library with mobile applications. *CMU Sch. Comput. Sci.* **2016**, *6*, 20.
24. Fang, M.; Boutros, F.; Damer, N. Unsupervised Face Morphing Attack Detection via Self-paced Anomaly Detection. *arXiv* **2022**, arXiv:2208.05787.
25. Neto, P.C.; Boutros, F.; Pinto, J.R.; Damer, N.; Sequeira, A.F.; Cardoso, J.S.; Bengherabi, M.; Bousnat, A.; Boucheta, S.; Menotti, D.; et al. OCFR 2022: Competition on Occluded Face Recognition from Synthetically Generated Structure-Aware Occlusions. *arXiv* **2022**, arXiv:2208.02760.
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
27. Thakur, N.; Han, C.Y. Indoor Localization for Personalized Ambient Assisted Living of Multiple Users in Multi-Floor Smart Environments. *Big Data Cogn. Comput.* **2021**, *5*, 42. [CrossRef]
28. Guerra, B.M.V.; Schmid, M.; Beltrami, G.; Ramat, S. Neural Networks for Automatic Posture Recognition in Ambient-Assisted Living. *Sensors* **2022**, *22*, 2609. [CrossRef]
29. Zhao, G.; Pietikainen, M. Dynamic Texture Recognition using Local Binary Patterns with an Application to Facial Expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 915–928. [CrossRef]
30. Zhong, L.; Liu, Q.; Yang, P.; Liu, B.; Huang, J.; Metaxas, D. Learning active facial patches for expression analysis. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2562–2569.

31. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE computer society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–26 June 2005; pp. 886–893.
32. Shan, C.; Gong, S.; McOwan, P.W. Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image Vis. Comput.* **2009**, *27*, 803–816. [CrossRef]
33. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
34. Berretti, S.; Ben Amor, B.; Daoudi, M.; DEL Bimbo, A. 3D facial expression recognition using SIFT descriptors of automatically detected keypoints. *Vis. Comput.* **2011**, *27*, 1021–1036. [CrossRef]
35. Carcagnì, P.; Del Coco, M.; Leo, M.; Distante, C. Facial expression recognition and histograms of oriented gradients: A comprehensive study. *SpringerPlus* **2015**, *4*, 645. [CrossRef]
36. Shan, C.; Gong, S.; McOwan, P. Robust facial expression recognition using local binary patterns. In Proceedings of the IEEE International Conference on Image Processing 2005, Genoa, Italy, 11–14 September 2005; Volume 2. [CrossRef]
37. Thakur, N.; Han, C.Y. Country-Specific Interests towards Fall Detection from 2004–2021: An Open Access Dataset and Research Questions. *Data* **2021**, *6*, 92. [CrossRef]
38. Wang, Z.; Wang, G.; Huang, B.; Xiong, Z.; Hong, Q.; Wu, H.; Yi, P.; Jiang, K.; Wang, N.; Pei, Y.; et al. Masked face recognition dataset and application. *arXiv* **2020**, arXiv:2003.09093.
39. Yovel, G.; Duchaine, B. Specialized Face Perception Mechanisms Extract Both Part and Spacing Information: Evidence from Developmental Prosopagnosia. *J. Cogn. Neurosci.* **2006**, *18*, 580–593. [CrossRef]
40. Ekman, P. Facial expression and emotion. *Am. Psychol.* **1993**, *48*, 384. [CrossRef]
41. Kotsia, I.; Buciu, I.; Pitas, I. An analysis of facial expression recognition under partial facial image occlusion. *Image Vis. Comput.* **2008**, *26*, 1052–1067. [CrossRef]
42. Ciregan, D.; Meier, U.; Schmidhuber, J. Multi-Column deep neural networks for image classification. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3642–3649.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 630–645.
44. Jin, H.; Liao, S.; Shao, L. Pixel-in-Pixel Net: Towards Efficient Facial Landmark Detection in the Wild. *Int. J. Comput. Vis.* **2021**, *129*, 3174–3194. [CrossRef]
45. Li, H.; Wang, N.; Ding, X.; Yang, X.; Gao, X. Adaptively Learning Facial Expression Representation via C-F Labels and Distillation. *IEEE Trans. Image Process.* **2021**, *30*, 2016–2028. [CrossRef]
46. Bargal, S.A.; Barsoum, E.; Ferrer, C.C.; Zhang, C. Emotion recognition in the wild from videos using images. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; pp. 433–436.
47. Zhang, J.; Kan, M.; Shan, S.; Chen, X. Occlusion-Free Face Alignment: Deep Regression Networks Coupled with De-Corrupt AutoEncoders. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3428–3437. [CrossRef]
48. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 17–24 May 2018; pp. 3–19.
49. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Chintala, S.; et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Processing Syst.* **2019**, 32.
50. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Zheng, X.; et al. Tensorflow: Large-Scale machine learning on heterogeneous distributed systems. *arXiv* **2016**, arXiv:1603.04467.
51. Croci, M.L.; Sengupta, U.; Juniper, M.P. Online parameter inference for the simulation of a Bunsen flame using heteroscedastic Bayesian neural network ensembles. *arXiv* **2021**, arXiv:2104.13201.
52. Qureshi, A.S.; Roos, T. Transfer Learning with Ensembles of Deep Neural Networks for Skin Cancer Detection in Imbalanced Data Sets. *arXiv* **2021**, arXiv:2103.12068.
53. Jain, A.; Kumar, A.; Susan, S. Evaluating Deep Neural Network Ensembles by Majority Voting Cum Meta-Learning Scheme. In *Soft Computing and Signal Processing*; Springer: Singapore, 2021; Volume 410, pp. 29–37. [CrossRef]
54. Liu, M.; Li, S.; Shan, S.; Chen, X. Au-Aware deep networks for facial expression recognition. In Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, 22–26 April 2013; pp. 1–6.
55. Zeng, N.; Zhang, H.; Song, B.; Liu, W.; Li, Y.; Dobaie, A.M. Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing* **2018**, *273*, 643–649. [CrossRef]
56. Ding, H.; Zhou, S.K.; Chellappa, R. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 118–126.
57. Miao, S.; Xu, H.; Han, Z.; Zhu, Y. Recognizing Facial Expressions Using a Shallow Convolutional Neural Network. *IEEE Access* **2019**, *7*, 78000–78011. [CrossRef]
58. Barros, P.; Churamani, N.; Sciutti, A. The FaceChannel: A Fast and Furious Deep Neural Network for Facial Expression Recognition. *SN Comput. Sci.* **2020**, *1*, 321. [CrossRef]
59. Li, M.; Xu, H.; Huang, X.; Song, Z.; Liu, X.; Li, X. Facial Expression Recognition with Identity and Emotion Joint Learning. *IEEE Trans. Affect. Comput.* **2018**, *12*, 544–550. [CrossRef]