symmetry

MDPI

Article

A Triple-Structure Network Model Based upon MobileNet V1 and Multi-Loss Function for Facial Expression Recognition

Baojin Han [1,2,*], Min Hu [1,2], Xiaohua Wang [1,2] and Fuji Ren [2,3]

[1] Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education, Hefei University of Technology, Hefei 230601, China
[2] School of Computer and Information, Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine, Hefei University of Technology, Hefei 230601, China
[3] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610056, China
* Correspondence: hanbaojin@mail.hfut.edu.cn

Abstract: Existing facial expression recognition methods have some drawbacks. For example, it becomes difficult for network learning on cross-dataset facial expressions, multi-region learning on an image did not extract the overall image information, and a frequency multiplication network did not take into account the inter-class and intra-class features in image classification. In order to deal with the above problems, in our current research, we raise a symmetric mode to extract the inter-class features and intra-class diversity features, and then propose a triple-structure network model based upon MobileNet V1, which is trained via a new multi-branch loss function. Such a proposed network consists of triple structures, viz., a global branch network, an attention mechanism branch network, and a diversified feature learning branch network. To begin with, the global branch network is used to extract the global features of the facial expression images. Furthermore, an attention mechanism branch network concentrates to extract inter-class features. In addition, the diversified feature learning branch network is utilized to extract intra-class diverse features. The network training is performed by using multiple loss functions to decrease intra-class differences and inter-class similarities. Finally, through ablation experiments and visualization, the intrinsic mechanism of our triple-structure network model is proved to be very reasonable. Experiments on the KDEF, MMI, and CK+ datasets show that the accuracy of facial expression recognition using the proposed model is 1.224%, 13.051%, and 3.085% higher than that using MC-loss (VGG16), respectively. In addition, related comparison tests and analyses proved that our raised triple-structure network model reaches better performance than dozens of state-of-the-art methods.

Keywords: facial expression recognition; MobileNet V1; symmetry and asymmetry; machine learning; deep learning; attention mechanism

Citation: Han, B.; Hu, M.; Wang, X.; Ren, F. A Triple-Structure Network Model Based upon MobileNet V1 and Multi-Loss Function for Facial Expression Recognition. *Symmetry* **2022**, *14*, 2055. https://doi.org/10.3390/sym14102055

Academic Editor: José Carlos R. Alcantud

Received: 30 August 2022
Accepted: 27 September 2022
Published: 2 October 2022

1. Introduction

With the progress of technology, human–machine communication has been merged into our lives [1–4]. The applications of facial expression recognition (FER) have become ever more essential, such as human–computer interaction, online testing, medical care, etc. [5–7]. When applying a facial expression classification system, the various expressions have serious differences. That can be demonstrated by the FACS (Facial Action Coding System) [8]. AU (action unit) [9] and expressions have some correspondence. For example, there is a greater symmetrical similarity between happiness and contempt compared with the symmetrical similarity between happiness and sadness because happiness and contempt contain AU12, and there is no intersection between the AU domain of happiness and sadness. Similar facial expressions can sometimes make it difficult to make a distinction.

Computer recognition of facial expressions mainly consists of three steps, viz., image preprocessing, feature extraction, and classification. Among them, feature extraction is

an important step. It relates to the recognition accuracy of facial expressions. Traditional feature extraction methods are mainly designed manually, such as Gabor wavelet, local binary pattern (LBP), histogram of gradient (HOG), etc. Rujirakul et al. [10] proposed a facial expression recognition method that contained histogram equalization (HE), principal component analysis (PCA), and extreme learning machine (ELM). HE was utilized for preprocessing to adjust the histogram curve of the input image. Then, PCA was employed to extract the features. Finally, ELM was employed for classification. Kumary et al. [11] put forward a facial expression recognition system which was the feature selection approach from the quantum-inspired binary gravitational search algorithm (QIBGSA). The idea of the QIBGSA was a modified binary version of the gravitational search algorithm by impersonating the properties of quantum mechanics. The experiment has achieved certain results. Islam et al. [12] presented a framework for recognizing human emotion through facial expression recognition by analyzing a large number of facial expression images and the possible locations of the expression regions in these images to manually segment the expression regions in an efficient and unique way. The experiment obtained better results. Xi et al. [13] raised surface electromyography (sEMG)-based emotion distribution learning (EDL) for predicting the intensity of underlying emotions. Choudhary et al. [14] proposed a systematic comparison of the facial features. Traditional methods in facial expression recognition applications can be found in [15–18]. Traditional feature extraction methods have many drawbacks, such as incomplete and limited information extraction, and insufficient robustness of image size and illumination.

As computer software and hardware evolve, environments for deep learning are well developed. The advantage of the convolutional neural network (CNN) is remarkable. The CNN can extract the features of images more completely and has strong robustness to the size and illumination of the images. It also has achieved good results in facial expression recognition. AlexNet [19,20], VGGNet [21,22], GoogLeNet [23], etc., have been commonly used for facial expression recognition. Due to the poor effect of traditional methods of facial expression recognition, Wu et al. [24] optimized and improved the internal structure based on LeNet-5 network. Batch normalization had been added to settle the over-fitting issue of the network owing to distinct features. Maximum pooling and average pooling were symmetrically used to fully extract facial expression features and to reduce the redundant data. Using deep learning to recognize facial expressions can enable the learning of important and robust features for different samples. This is a key problem with facial expression recognition. Ye et al. [25] presented a region-based convolutional fusion network (RCFN) to solve the problems by three aspects, which were a built muscle movement model, a constructed network, and constrained punitive loss. The experiment results showed that RCFN was effective in commonly used datasets. Singh et al. [26] proposed the classification of FER which used CNNs based on static images. Feature extraction was used to extract features of the facial part, such as eyes, nose, and eyebrows. The experiment achieved better results. Chen et al. [27] put forward an improved method of facial expression recognition based on CNN. A new convolution neural network structure was designed which uses a convolution kernel to extract key features and max pooling to reduce the redundant features. There are also deep learning methods in facial expression recognition applications, such as [28–37].

Many deep learning methods brought excellent results in cases of large amounts of computation, limiting their applications for small devices or offline scenarios. To effectively address this problem, Zhou et al. [38] proposed a frequency multiplication network (FMN), which was a deep learning method running in the frequency domain and could significantly reduce network capacity and computing workload. Combined with the uniform rectangle feature (URF), this method further improves the performance and reduces the training workload. Cotter et al. [39] put forward a new lightweight deep learning model, Mobi-ExpressNet for FER. The model relied on depthwise separable convolutions to limit the complexity and used a fast down sampling method and several layers in the architecture to keep the model size very small. It achieved good results. Nan et al. [40] proposed a

lightweight A-MobileNet model. The method significantly improved recognition accuracy without increasing the number of model parameters.

The continuous development of facial expression recognition technology has led to the continuous advancement of face detection and recognition technology. This promotes the development of facial expression recognition technology. Ding et al. [41] proposed a shared generative adversarial network, SharedGAN, to expand the gallery dataset. Experimental results illustrated the effectiveness of SharedGAN and showed satisfactory results. Abdulhussain et al. [42] presented a new scheme for face recognition which used hybrid orthogonal polynomials to extract features.

The above facial expression recognition methods have been improved in several aspects, but some problems still exist:

(1) The cross-dataset facial expression comes from different facial expression datasets with fuzziness and asymmetry, so differences among facial expressions are huge. It becomes more difficult for network learning on cross-dataset facial expressions, which results in a decrease in recognition accuracy.
(2) Multi-region learning on an image does not extract the overall image information. The corresponding network lacks global information. So, it also makes identification more difficult.
(3) A frequency multiplication network could reduce the network complexity, but it does not take into account the inter-class and intra-class features in image classification. This results in a low facial expression recognition rate.

Focusing on the above problems, we put forward a symmetric mode to extract the inter-class features and intra-class diversity features, and then propose a triple-structure network model, which is trained via a new multi-branch loss function. The proposed network consists of triple structures, i.e., a global branch network, an attention mechanism branch network, and a diversified feature learning branch network. The proposed network is based upon MobileNet V1, which has the characteristics of being lightweight and a high recognition rate. The focus is different from each branch loss function. The global branch network mainly focuses on learning the global features of images. The attention mechanism branch network mainly concentrates on learning the inter-class features, and the diversified feature learning branch network mainly focuses on learning the intra-class diversity features.

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

(1) A facial expression recognition network is proposed based upon MobileNet V1. Our network is a simple and effective network, which can achieve a better recognition rate.
(2) We propose an improved multi-loss function network, which includes a global branch network, an attention mechanism branch network based on SENet, and a diversified feature learning branch network. The global branch network is employed to extract the global features of facial expression images. A symmetric mode is raised to extract the inter-class key features and intra-class diversity features. In detail, the attention mechanism branch network concentrates to extract inter-class key features, while the diversified feature learning branch network is used to extract intra-class diverse features.
(3) We put forward a multi-branch network. The network avoids only focusing on the local or global regions of the image, but both global and local images participate in the learning.

The remainder of this study is organized as described next. Section 2 introduces related works, including MobileNet V1 and SENet. Section 3 depicts the details of our proposed network of triple structures. In Section 4, the effectiveness of our network is verified by comparison and ablation experiments with some facial expression recognition models, respectively. In Section 5, to further test the effectiveness of our network, a class activation map visualization is performed on each branch of the model. Section 6 summarizes and discusses future research plans.

2. Related Works

In this section, we slightly look back at the previous related research about MobileNet V1 and SENet and highlight their merits and achievements.

2.1. MobileNet V1

When performing facial expression recognition, so as to obtain a certain effect, some complex networks such as AlexNet, VGGNet, and GoogLeNet are often used. However, the complex networks will influence the magnitude and computational speed of the network. For example, when the complex networks were employed in automatic detection, the real-time nature of visual tasks and other factors need to be considered by reason of the limitations such as the computational speed on a platform. MobileNet V1 employs a simple and effective architecture of hyperparameters, which can enable fewer network parameters and speed up computation. Additionally, the network is very practicable for facial expression recognition. Figure 1 shows the MobileNet V1 network structure.

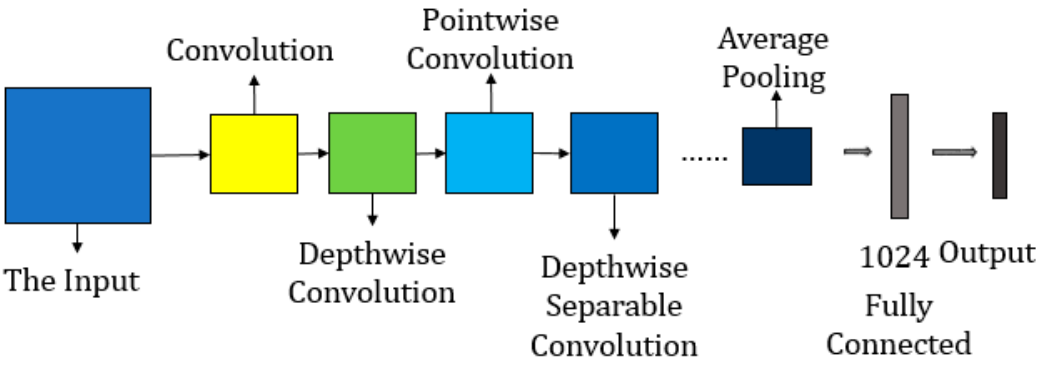

Figure 1. The structure of the MobileNet V1 network.

The core layer of MobileNet V1 is a deep separable filter. Deep separable convolution is a form of deconvolution. The standard convolution operation is divided into two steps: (1) extracting feature maps, (2) superimposing the extracted feature maps. The depth separable convolution separates the course of two layers. Firstly, one layer is the depthwise convolution, which is employed to extract features for each channel. Lastly, the other layer is a point-by-point convolution, which employs a 1×1 convolution to integrate the output of the first step. This decomposition is an uncomplicated and valid method that can significantly reduce superfluous calculations and optimize the network structure. Diagrams of standard convolution, depthwise convolution, and pointwise convolution are shown Figures 2–4, respectively. The size of the input feature map is $D_F \times D_F \times M$. Here, M stands for the number of input channels, D_F represents the size of the feature map, and N denotes the number of output channels, and the parameters of a standard convolutional layer show $D_k \times D_k \times M \times N$, and D_k indicates the size of the convolution kernel. If the size of the output feature map does not change, the computing costs of the standard convolution are

$$C = D_k \times D_k \times M \times N \times D_F \times D_F \quad (1)$$

The MobileNet V1 model employs deep separable convolutions to shatter the interaction between the number of output channels and the size of the kernel to effectively decrease redundant computing. The computing costs of depthwise convolution are

$$C_1 = D_k \times D_k \times M \times D_F \times D_F \quad (2)$$

Depthwise convolution which only filters the input channels does not generate new features, and an additional pointwise convolution is required to integrate the features gained by output filters to comprise new multi-channel features. So, depthwise convolution

is more quick than standard convolution. The computing costs of pointwise convolution C_2 are characterized by

$$C_2 = M \times N \times D_F \times D_F \tag{3}$$

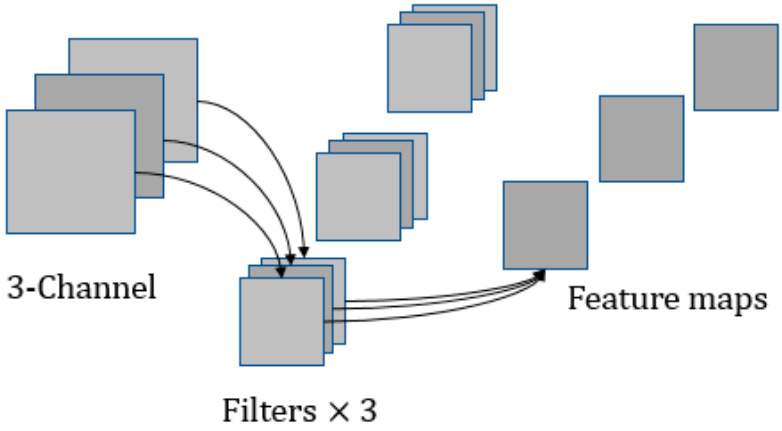

Figure 2. The structure diagram of standard convolution.

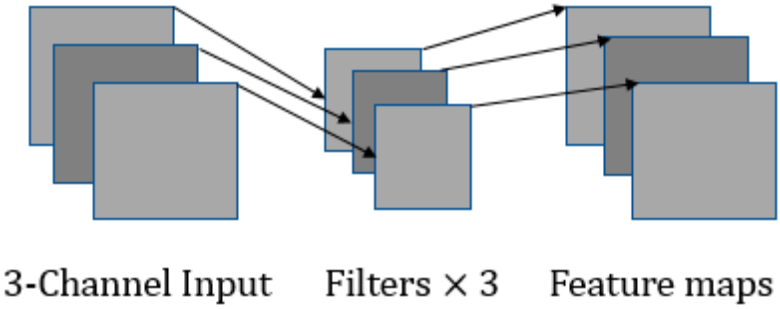

Figure 3. The structure diagram of depthwise convolution.

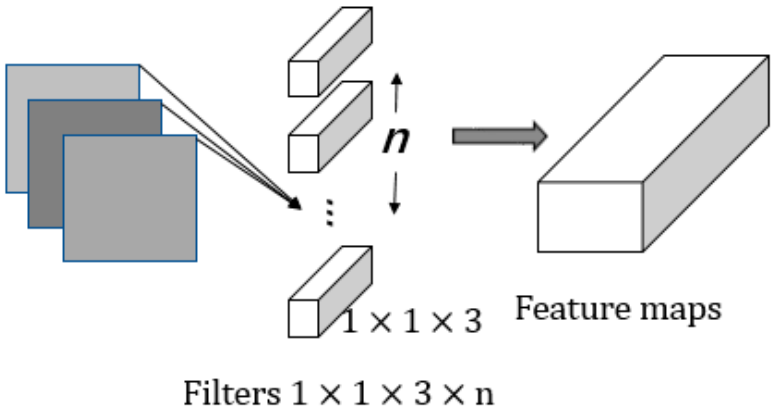

Figure 4. The structure diagram of pointwise convolution.

So as to decrease computing resources, the standard convolution integral is decomposed into depthwise convolution and pointwise convolution, as shown by

$$\frac{C_1 + C_2}{C} = \frac{D_k \times D_k \times M \times D_F \times D_F + M \times N \times D_F \times D_F}{D_k \times D_k \times M \times N \times D_F \times D_F} = \frac{1}{N} + \frac{1}{D_k^2} \tag{4}$$

If the relatively large value of N is considered, then assuming a 5×5 convolution kernel, the depthwise separable convolution is 25 times smaller than the standard convolution in terms of computational complexity. One step can show that compared with the standard

convolutional neural network, MobileNet V1 significantly decreases computing costs and narrow model size, enhancing computing speed of the model.

To resolve the question of facial expression recognition (FER) technology on masked faces, Yang et al. [43] proposed a method which could add face masks to existing FER datasets automatically. The results were feasible for the method.

Sadik et al. [44] improved the MobileNet model and implemented transfer learning technique. The outcome was satisfactory. Petrosiuk et al. [45] proposed a method which solved the problem of insufficient data volume in sets of images with different facial expressions. The developed technology of transfer learning of MobileNet and the subsequent "fine tuning" of the network parameters has led to new developments. Related experiments obtained good results.

It can be obtained from the above formula derivation and application examples, under the premise of the same feature map, that the computation and parameters of depthwise separable convolutions are greatly reduced. The MobileNet V1 network is a simple and effective network. Therefore, in this study, we chose MobileNet V1, which can raise the recognition rate of facial expression.

2.2. SENet

The SENet (squeeze-and-excitation networks) was first proposed by Hu et al. [46] and won the champion of the image classification task in the 2017 ILSVRC challenge. SENet is an attention mechanism that can be seamlessly merged into the CNN architecture with low overheads.

The core of the model is to let the MobileNet V1 learn different weights for different features, where the SENet can extract more corresponding facial expression features. The features that are not concerning to the facial expression are suppressed, and the learning between feature channels is reinforced to obtain more facial expression feature information, significantly ameliorating the facial expression recognition rate of the network. Figure 5 is a structural diagram of SENet, which largely consists of two parts: squeeze and excitation. These two parts finished the suited calibration of the feature channel.

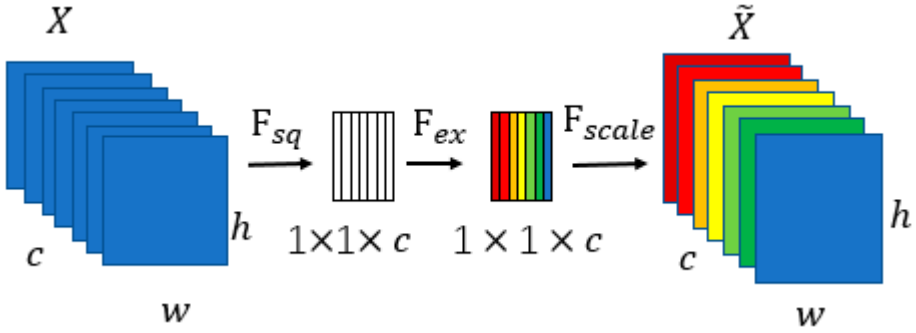

Figure 5. SENet network structure diagram.

The squeeze and excitation module promotes the network denotation capability by exploiting channel dependencies. The particulars of the SENet are displayed with Figure 5. Through the module, significant features are stressed on the channels while inhibiting redundant features. The squeeze function of Figure 5 is expressed as follows:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j), \tag{5}$$

where z_c is the c-th element of the squeezed channels and F_{sq} represents the squeeze function. u_c is the c-th channel of the input. H and W stand the height and width of the

input images. Then, an excitation function tacks the squeeze operation, which aims to fully seize the channel-wise interrelation. The excitation displays as below:

$$s_c = F_{ex}(z, W) = \sigma(W_u \delta(W_d z_c)) \tag{6}$$

in which F_{ex} stands for the excitation function and z_c is the input squeezed signal from the last layer. δ represents the ReLU activation function, $W_d \in \mathbb{R}^{c \times \frac{c}{r}}$ is the channel using the 1×1 kernel size and the dimensionality reduction ratio r to scale down. $W_u \in \mathbb{R}^{\frac{c}{r} \times c}$ is the channel with the ratio of r after being activated by ReLU. The final output of the block $\widetilde{x_c}$ is rescaled using the channel s_c shown below:

$$\widetilde{x_c} = F_{scale}(u_c, s_c) = s_c \cdot u_c \tag{7}$$

For the feature map ($X \in R^{w \times h \times c}$) input to the SENet, the output feature ($\widetilde{X} \in R^{w \times h \times c}$) was passed to the squeeze function (Figure 5 F_{sq}) and the excitation function (Figure 5 F_{ex}). The squeeze function was used to embed information from the global receptor into the channel descriptor in each layer. The squeeze function generated a sequence in $1 \times 1 \times c$, which indicated the interrelation among each layer. The excitation function later was used to carry out feature recalibration through reweighting the original feature mappings. So as to decrease the model parameters and preserve the high FER accuracy, Zhong et al. [47] put forward a simple and effective network based on squeeze-and-excitation (SENet) and ResNet. Beside the state-of-the-art methods using the visual geometry group (VGG) or other networks, the model improved the accuracy and reduced the model size, which was rival according to model size and recognition rate.

To resolve the problem of emotional recognition of speech, Zhao et al. [48] used parallel convolutional layers (PCN) integrated with the squeeze-and-excitation network (SENet) to extract relationships from 3D spectrograms across time steps and frequencies. The experiment achieved good results.

From the above formula derivation and application examples, it can be concluded that using SENet under the same conditions can better learn important features and suppress redundant features. Thus, the learned features are more important, which is beneficial to better enhance the facial expression recognition rate. Therefore, the paper chose to employ SENet, which can validly improve the recognition rate.

Table 1 displays a summary of the relevant works.

Table 1. A summary of the related works.

	Core Component	Advantage
MobileNet V1	A deep separable filter	Light and better recognition rate
SENet	Squeeze and excitation	Low overhead, integrated into the CNN architecture

3. The Proposed Triple-Structure Network Model

In this study, we propose a symmetric mode to extract the inter-class features and intra-class diversity features, and then put forward a triple-structure network model based on MobileNet V1, which is trained via a new multi-branch loss function. Such a proposed network consists of triple structures, which incorporates a global branch network, an attention mechanism branch network, and a diversified feature learning branch network. The overall architecture model of the proposed triple-structure network model is shown in Figure 6.

In what follows, we show the details for the proposed triple-structure network model.

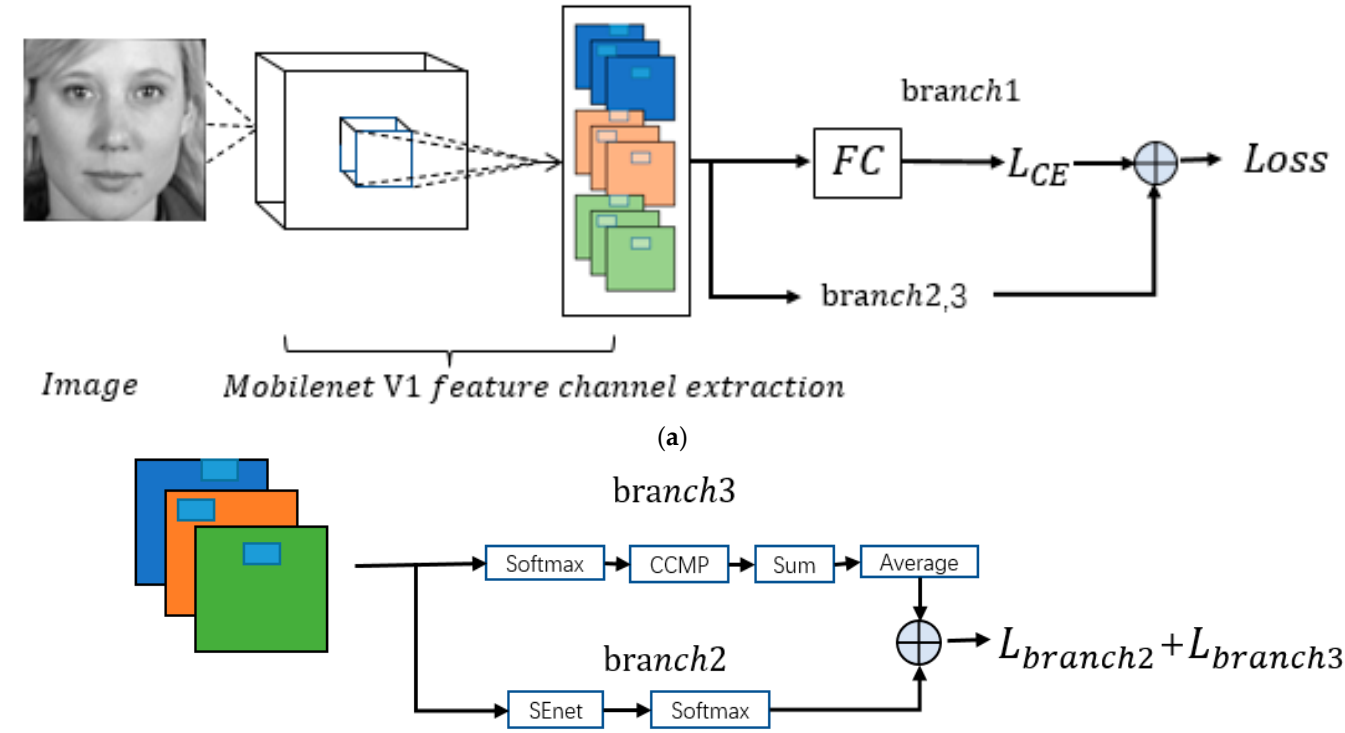

Figure 6. The idea of the proposed triple-structure network model: (**a**) the whole structure; (**b**) branch2 and branch3.

3.1. The General Idea of the Triple-Structure Network Model

Inspired by Ref. [49], our network introduces mutual channel loss (MC-loss) to discover multiple discriminative regions and restrict feature distribution. MC-loss has the following advantages:

- It makes the network easier to train, because the network does not introduce any additional network parameters.
- The method combines global and local regions and forces the network to capture subtle differences by discriminant components.
- It can effectively improve the recognition rate for solving fine-grained image classification.

In the meantime, MC-loss avoids too much attention to localized regions of an image. Thus, a global branch has been introduced to extract information from the overall regions of an image.

Based on the above advantages, our study applies MC-loss to facial expression recognition. The network is adjusted according to the differences between the expression image set and the fine-grained image set. For example, fine-grained image classification emphasizes the problem of distinguishing among subcategories of common visual categories. Meanwhile, facial expression recognition emphasizes the classification of facial motion.

MobileNet V1 is used instead of the ResNet/VGG on MC-loss. Additionally, MC-loss has utilized the discriminative component to promote network learning. However, the discriminative component contains a random attention mechanism, such as the feature F_i corresponding to each class of the feature F extracted by the basic network is randomly discarded by the half. This random discarding method influences the recognition accuracy. Our network employs the attention mechanism branch network to take the place of the discriminative component. The attention mechanism branch does not contain the random

discarding method. This branch can effectively distinguish the inter-class features of facial expression images.

The overall framework is displayed in Figure 6. Based on the basic backbone network of MobileNet V1, our network consists of triple structures (or branches), i.e., the global branch network (branch1), the attention mechanism branch network (branch2), and the diversified feature learning branch network (branch3).

Branch1 learns mainly the global features. Branch2 acquires primarily to extract inter-class key features, and branch3 focuses to extract intra-class diverse features. Note that branch2 and the branch3 construct a symmetric mode to extract the inter-class features and intra-class diversity features.

The total network loss function *Loss* can be defined as follows:

$$Loss = L_{branch1} + \mu L_{branch2} + \lambda L_{branch3}. \tag{8}$$

$L_{branch1}$ represents the branch1 loss function. $L_{branch2}$ stands for the branch2 loss function. $L_{branch3}$ denotes the branch3 loss function. μ and λ are hyper-parameters. During the training phase, the global network branch (branch1) can fully learn the inter-class and intra-class features in facial expression images through the guidance of branch2 and branch3. Thus, branch2 and branch3 do not participate in the test phase.

3.2. Basic Backbone Network and Global Branch Network

In our research, to enter an image, we first extract the feature map by feeding the image into MobileNet V1. The extracted feature maps are represented as $\mathcal{F} \in R^{H \times W \times N}$. H stands for the height, W represents the width, and N is the number of channels.

Additionally, we need to set the value of N to be equal to $C \times \xi$. C is the number of classes in a dataset. ξ stands for the number of feature channels employed to express each class.

Therefore, the n-th vectored feature channel of $\mathcal{F}$ can be expressed by

$$\mathcal{F}_n \in R^{WH}, \quad n = 1, 2, \ldots, N. \tag{9}$$

Please notice that we reshaped each channel matrix of $\mathcal{F}$ of dimension $W \times H$ to a vector of size W multiplied by H, i.e., WH. Therefore, the grouped feature channels relevant to the i-th class is pointed to by $F_i \in R^{WH}$, where $i = 0, 1, 2, \ldots, C-1$, namely:

$$F_i = \left\{ \mathcal{F}_{i \times \xi + 1}, \mathcal{F}_{i \times \xi + 2}, \ldots, \mathcal{F}_{i \times \xi + \xi} \right\}. \tag{10}$$

As a consequence, we obtained the grouped deep features F, in which $F = \{F_0, F_1, \ldots, F_{c-1}\}$. $g(F_i)$ represents the F_i processed by the fully connected layer. We made use of the cross-entropy loss function L_{CE}, calculating the dissimilarity between the ground-truth label y and the predicted probability *pred*. Here, *pred* is expressed by the following form:

$$pred = \frac{\left[e^{g(F_0)}, e^{g(F_1)} \ldots, e^{g(F_{c-1})} \right]^T}{\sum_{i=0}^{c-1} e^{g(F_i)}} \tag{11}$$

To sum up, the loss function of branch1 (i.e., $L_{branch1}$) can be described as below:

$$L_{branch1} = L_{CE}(y, pred) \tag{12}$$

3.3. Attention Mechanism Branch Network

In facial expression recognition, because muscle activity is very similar to intra-class facial expression, the intra-class similarity is high. Distinguishing different facial expressions is a significant step in recognition that directly affects the accuracy of recognition. Feature extraction is carried out for facial expression images, while differentiated weighting is performed on the extracted features. Features processed by SENet can improve the discrepant features, and thus better distinguish different emotions.

For this reason, we employed SENet to assign weights to facial expression features, which makes the important feature channels play a bigger role and weakens the unimportant feature channels. Through the SENet operation, the weight of each channel is different, which makes it easier to distinguish efficiently different facial expressions. We define $pred_{Senet}$ as follows:

$$pred_{Senet} = \frac{\left[e^{g_{SENet}(F_0)}, e^{g_{SENet}(F_1)}, \ldots, e^{g_{SENet}(F_{c-1})}\right]^T}{\sum_{i=0}^{c-1} e^{g_{SENet}(F_i)}} \tag{13}$$

Here, $g_{SENet}(F_i)$ represents the feature F_i processed by SENet. We used the cross-entropy loss function L_{CE} to calculate the dissimilarity between the ground-truth label y and the predicted probability $pred_{SENet}$.

Following that, the loss function of branch2 (i.e., $L_{branch2}$) can be characterized by the following formula:

$$L_{branch2} = L_{CE}(y, pred_{SENet}). \tag{14}$$

3.4. Diversified Feature Learning Branch Network

Considering facial expression images, different feature channels of the same class should focus on dissimilar areas of the facial expression images, instead of the total of the channels concentrating on discriminative areas. For example, the global branch network and the attention mechanism branch network have difficulty in capturing different regions on the same facial expressions, resulting in a low recognition rate.

To better solve the above problem, a diversified feature learning branch network was introduced to our research. We attempted to learn multiple regions within a class and utilize multiple losses to supervise the training, which allows multiple regions to work commonly and symmetrically complement each other. The fundamental purpose of the diversity loss function is to yield the feature maps within a class, which is different regions during learning. Thus, the learned feature of the class is more diverse. The specific expression of the loss function of branch3 (i.e., $L_{branch3}$) is shown as follows:

$$L_{branch3} = \frac{1}{c} \sum_{i=0}^{c-1} h(F_i) \tag{15}$$

in which

$$h(F_i) = \sum_{k=1}^{WH} \underbrace{\underset{j=1,2,\ldots,\varepsilon}{max}}_{CCMP} \underbrace{\left[\frac{e^{F_{i,j,k}}}{\sum_{k'=1}^{WH} e^{F_{i,j,k'}}}\right]}_{Softmax} \tag{16}$$

Here, W expresses the width of the feature map, and H represents the height of the feature map. ε stands for the number of characteristic graphs in each intra-class of expressions. As shown in Figure 6, the feature map is normalized by softmax, and then dealt with by CCMP (cross channel max pooling). The CCMP comes from the concept of maxout [50], which takes out the channels of each class and their maximum values. Through the above, some distinguished features within a class can be concentrated on a one-dimensional feature map. Then, salient regions of each group of features are obtained by accumulation, which are averaged to obtain the feature diversity loss $L_{branch3}$.

4. Experiments and Analyses

4.1. Experimental Dataset

To evaluate the proposed triple-structure network model, we conducted experiments on the MMI [51], KDEF [52], and CK+ [53] databases. KDEF images are captured in controlled lab environments, containing 4900 images consisting of 70 people, of which 35 are females and 35 are males, aged between 20 to 30, which display 7 basic facial expressions.

We only employed the 980 front view images. MMI is also a lab-controlled database with six basic expressions. We selected the three peak pictures from each sequence. CK+ is also a lab-controlled database with seven basic expressions. As in the case of the MMI database, we chose also the three peak frames from each sequence. Figure 7 shows part of the processed images. The first three images are the processed pictures in the MMI, and the middle three pictures are the processed pictures in the KDEF dataset, and the last three pictures are the processed images in the CK+ dataset. Table 2 displays the number of different emotion pictures in KDEF, MMI, and CK+ datasets. In this study, we tested a total of 2570 images: 609 for the MMI dataset, 980 for the KDEF dataset, and 981 for CK+.

Figure 7. Part of the processed images.

Table 2. The number of different emoticon pictures in KDEF, MMI, and CK+ datasets.

Class	Happy	Anger	Sad	Disgust	Surprise	Fear	Neutral	Contempt
KDEF	140	140	140	140	140	140	140	0
MMI	42	32	32	28	41	28	0	0
CK+	69	45	28	59	83	25	0	18

In order to prevent over fitting and increase prediction robustness, we conducted data augmentation to the MMI, KDEF, and CK+ datasets. Specifically, we randomly created 10 cropped images of size 224 × 224 for the original images, whose sizes were all 240 × 240. Furthermore, we also collected 10 processed images for each facial expression to test by cropping the top left corner, bottom left corner, top right corner, bottom right corner, the center, and subsequently taking the reflection of each of these cropped images. We made the final decision by taking the average results of these 10 processed images to reduce the classification error.

4.2. Experimental Settings

The experiments were used in the environment of Python 3.6.10, pytorch 1.6.0, TensorFlow 1.14.0 and an operating system of 18.04.1-Ubuntu. The proposed triple-structure network model in the experiments was run on a computer with Intel(R) Xeon(R) CPU E5-2620v3@2.40GHz in CPU and two 12G Nvidia GeForce GTX1080Ti graphics cards in GPU. In the experiments, GPUs were used to speed up the model calculation and reduce the training time. The stochastic gradient descent (SGD) method of momentum parameters in small batches was selected as the model parameter optimizer. The learning rate of the network was initially set to 0.1. Between the 150th and 225th iterations, the rate was set to 0.01. Beyond 225 iterations, we set it to 0.001. μ and λ were set to 1.5 and 20, respectively. The rest of the relevant settings are shown in Table 3 below.

Table 3. Related parameter settings.

Items	Settings
SGD	0.9
Loss function	Cross Entropy Loss
Training period	300 times
Batch size	128

4.3. Ablation Experiments

For the sake of analysis, we performed an extensive ablation study by removing certain portions of our proposed triple-structure network model to see how that affected performance. This was carried out using the MMI, KDEF, and CK+ databases.

Our triple-structure network model is mainly composed of three branches, i.e., a global branch network (MobileNet), an attention branch network (SENet), and a diversified feature learning branch network (Div). To verify its effectiveness, we compared it with MobileNet, MobileNet+SENet, and MobileNet+Div, respectively. MobileNet stands for a global branch network. MobileNet+SENet represents the combination of the global branch network and attention mechanism branch network. Finally, MobileNet+Div denotes the combination of the global branch network and diversified feature learning branch network. Since our study is inspired by Ref. [49], the network of [49] was also used as a comparison. Each network was under the same setting as in Section 4.2. Table 4 displays the outcomes of comparative data. The recognition accuracy (%) was used for performance evaluation.

Table 4. Ablation studies for key modules of our triple-structure network model on the KDEF, MMI, and CK+ databases.

Branch	KDEF	MMI	CK+
MC-loss (VGG16) [49]	95.306	70.508	95.957
MobileNet	95.204	77.627	94.255
MobileNet+SENet	96.122	77.796	96.595
MobileNet+Div	95.510	79.322	95
Our network model	96.530	83.559	99.042

From the backbone network, MobileNet had a good recognition rate for facial expression images. In the case of the KDEF dataset, we can discover the advantages of our proposed triple-structure network model. Compared with the backbone network, the recognition rate of MobileNet+SENet was increased by nearly 1%. Because the KDEF dataset is mainly composed of images of young men and women aged 20 to 30, inter-class features have great differences. Therefore, the recognition effect was significantly improved. However, when MobileNet+Div learns intra-class features, the similarity of intra-class features was too high. Therefore, the recognition rate of MobileNet+SENet was better than that of MobileNet+Div. Our triple-structure network model combines the strengths of the MobileNet+SENet and the MobileNet+Div. Therefore, the recognition rate has been significantly improved.

In Table 4, the recognition rate of the MMI dataset was obviously not as high as the one in the KDEF dataset. This was due to the following reasons: the age, facial shape, facial occlusions, and so on, resulting in the overall recognition effect as not as ideal as the KDFE dataset. In the MMI dataset, compared with MobileNet, the improvement of the recognition rate of MobileNet+SENet was not obvious. This was due to the huge differences in facial expression images or face occlusions. Therefore, it is difficult to have further improvement. In the MobileNet+Div, it learned a wealth of intra-class features in the MMI dataset. Thus, the learnable range of intra-class features was increased, which makes a remarkable improvement. Our triple-structure network model combines the advantages of the MobileNet+SENet and the MobileNet+Div, so the overall improvement effect was better in the MMI dataset. The feature extraction network applied in Ref. [49] is VGG16 (with MC-loss (VGG16)). It can be discovered from the comparison of Table 4 that Ref. [49] has a little advantage in the KDEF dataset. This is because the facial expression differences in the KDEF dataset are small. The MMI dataset can reflect the benefits of our triple-structure network model, which shows that our triple-structure network model also has great advantages for expression images with large differences.

Since the expressions in the CK+ database have obvious characteristics and are mainly composed of young men and women, they have achieved good recognition results in the MobileNet. With the progress of learning, MobileNet+SENet had a good effect in

learning the inter-class features in the CK+ database, because the distinction of the inter-class features in the CK+ database was obvious, so the learning effect was good. Ref. [49] obtain better results in the CK+ database, but its effect is not so obvious compared with our network. Note that our network learns from multiple aspects, but Ref. [49] randomly throws away some inter-class features, resulting in its recognition effect being not as good as our network.

4.4. Comparison with Other Methods

Table 5 displays the performance comparisons with the same approaches. To summarize, our triple-structure network model obtained competitive results on KDEF.

Ref. [10] uses the traditional PCA method to extract features and employs ELM for classification, because the feature information extracted by CNN is richer than PCA. Therefore, the recognition rate of [10] is lower than that of our triple-structure network model. Ref. [11] makes use of LBP and Gabor filtering methods to extract features. Due to the limitations of traditional feature extraction methods, its recognition rate is not as high as our triple-structure network model. Ref. [12] learns the possible locations of expression regions in many images, but the expression regions vary widely across datasets. Recognized cross-image datasets tend to lead to more recognition errors and use traditional feature extraction methods, which is not an ideal result.

The muscle models proposed in Ref. [25] are segmented, the key region features are fused, and a penalty loss function is added simultaneously. These methods can enhance the expression recognition rate. However, there is also a gap compared to the global-to-local loss function in our triple-structure network model. Ref. [54] utilizes the FER network to effectively identify FER with the help of the softmax classifier. Since only a single network is used and there is no multi-angle learning, the recognition accuracy of the network of [54] is not so high as that of our triple-structure network model.

Table 5. The recognition accuracy (%) of the related methods in the KDEF datset.

Methods	KDEF
HE+DeepPCA+ELM [10]	83.00
QIBGSA [11]	92.35
Ref. [12]	86.84
RCFN [25]	91.60
FER-net [54]	83.00
Our network model	96.53

Table 6 shows the performance comparisons with the same approaches. To summarize, our triple-structure network models obtained competitive results from the MMI datasets.

Table 6. The recognition accuracy (%) of the related methods in the MMI dataset.

Methods	MMI
FMN [38]	81.39
DLP+CNN [55]	78.46
DeRL [56]	73.23
[57]	80.70
Our network	83.56

Ref. [38] shows good advantages but did not involve regional learning. Single global learning leads to a low recognition rate. By comparison, our triple-structure network model is more ideal. Ref. [55] proposes an expectation maximization algorithm that estimates emotion labels. This reveals those facial expressions of the real world that often express complex or even mixed emotions, and multi-label facial expressions often lead to false recognition in learning and result in low classification accuracy. Ref. [56]

employs de-expression residual learning (DeRL) to extract the information about expression components to recognize facial expressions. Since neutral expressions are generated in the generative model, it will inevitably result in the deformation of some expressions. Thus, the recognition rate is not as good as that of our triple-structure network model which used the original images directly. Ref. [57] learns the spatial features and temporal dynamics of FER, but the corresponding temporal and spatial dynamics learning results in many learning features, which are causing redundancy. Overall, our triple-structure network model performs the best.

Table 7 shows the performance comparisons with the same approaches. To summarize, our triple-structure network model obtained competitive results on the CK+ datasets.

Table 7. The recognition accuracy (%) of the related methods in the CK+ dataset.

Methods	CK+
MF-MLP [58]	98.06
Ref. [59]	98.01
HAAR+LDA+IBH-based ELM [60]	97.62
HAAR+PCA+IBH-based ELM [60]	96.27
Ref. [61]	94.8
SGWT [62]	96.15
Our network	99.042

Ref. [58] proposed a multi-feature-based MLP (MF-MLP) classifier which is focused on the facial appearance detection problem. MF-MLP used LBP to extract features, but traditional feature extraction methods have some limitations. Therefore, the performance of our network is better than that of MF-MLP. Ref. [59] presented a smile classification method which is based on an association of row transform-based feature extraction algorithm and ensemble classifier. The methods lacks global features, so the effect is not as good as that of our network. Ref. [60] used the LDA and PCA to decrease the dimensions of the face images and maintain the most important features. The effective information extracted by LDA and PCA is not as comprehensive and robust as CNN, and the experimental effect is slightly worse than our network. Ref. [61] used the VGG16 model for a modified trained model and attaches additional layers on it. However, the model lacks local information, and so our network has a certain advantage. Ref. [62] leveraged spectral graph wavelet transform to extract information. Mobile V1 can automatically find key feature information, while spectral graph wavelet transform can only be used to extract information (containing redundant information). Therefore, our network performs better than spectral graph wavelet transform.

In [10–12,58,60], they all used the traditional feature extraction methods. Due to the traditional methods having their limitations in feature extraction, their recognition rate is not as high as that of our triple-structure network model. This again verifies that the traditional feature extraction method does have certain limitations for the feature information extraction of images. Comparing the results of the method in this paper with [25,38,54,59,61], it can be concluded that single local or global learning is not as good as joint local and global learning. The images used in Refs. [55,56] contain multiple expressions, extracting many redundant features, resulting in a low recognition rate.

According to the experimental results and analyses, the proposed triple-structure network model is more competitive than other methods.

4.5. Confusion Matrices and the Classification Report

In Figures 8–19, Hap represents happy, Ang is anger, and Sad stands for sad. Dis means disgust, Sup is surprise, and Fea denotes fear. Neu means neutral and Con is contempt. The precision is the proportion of correctly predicted data in each class to all data of that class. The recall rate (recall) reflects the proportion of correctly predicted data

in each class to all data predicted to be of that class. F1 reflects the performance of the model by combining the two indicators of precision and recall, as shown below:

$$F1 = 2 \times \frac{P + R}{P \times R} \tag{17}$$

In (17), P represents precision and R is recall. TR is total recall, and TP is total precision, and $TF1$ is total $F1$-score.

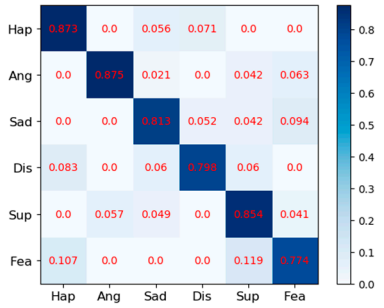

Figure 8. Performance of confusion matrix on the MMI dataset.

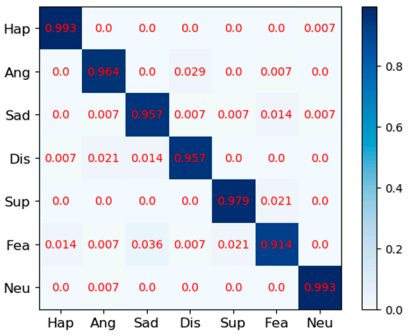

Figure 9. Performance of confusion matrix on the KDEF dataset.

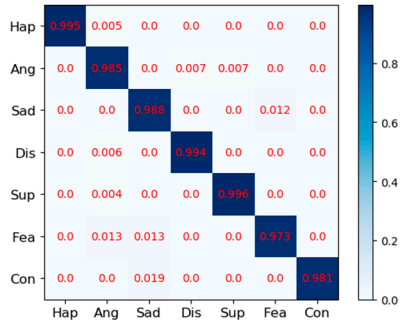

Figure 10. Performance of confusion matrix on the CK+ dataset.

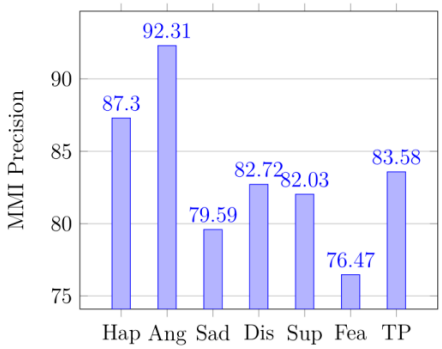

Figure 11. Performance of precision report on the MMI database.

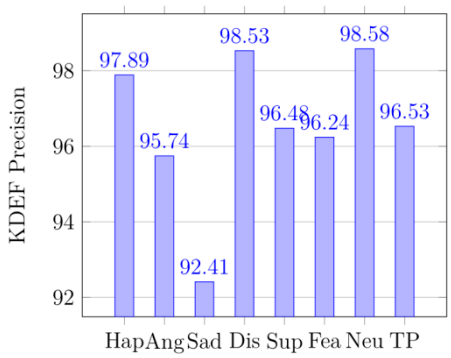

Figure 12. Performance of precision report on the KDEF database.

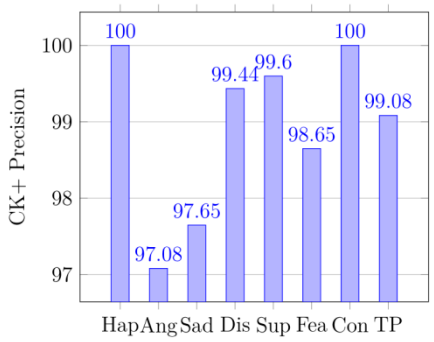

Figure 13. Performance of precision report on the CK+ database.

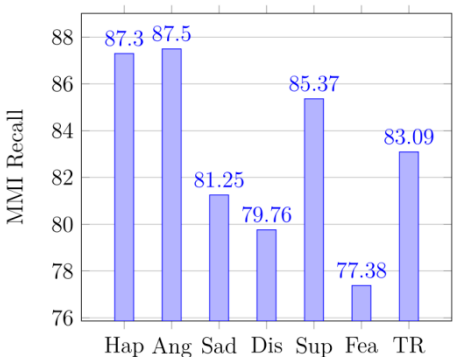

Figure 14. Performance of recall report on the MMI database.

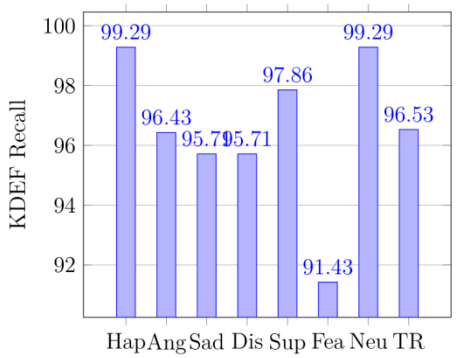

Figure 15. Performance of recall report on the KDEF database.

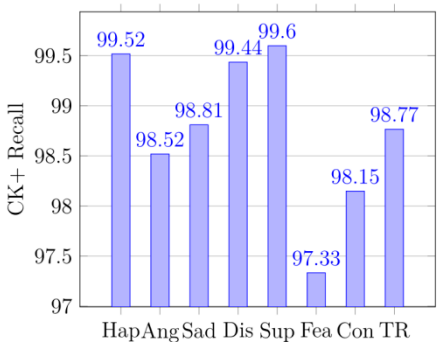

Figure 16. Performance of recall report on the CK+ database.

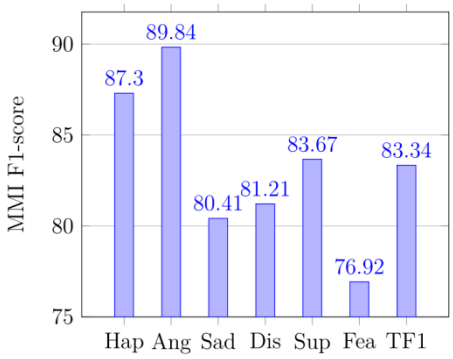

Figure 17. Performance of F1-score report on the MMI database.

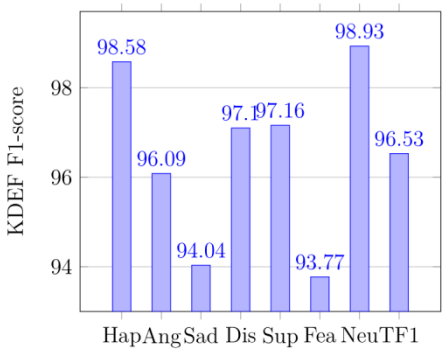

Figure 18. Performance of F1-score report on the KDEF database.

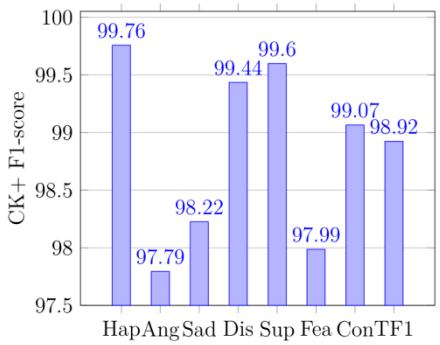

Figure 19. Performance of F1-score report on the CK+ database.

In the task, four popularly used evaluation metrics, namely, accuracy, precision, recall, and F1-score [63–66], were deliberated in the research to contrast the performance of our network to 21 models. However, the values of all the metrics were computed from the

confusion matrix. Confusion matrices can compute the values of metrics, which were employed further to dissect the function of our network. Figure 12 displays the confusion matrix of the KDEF dataset. The size of the confusion matrix is seven by seven as seven facial expressions were premeditated for this research. The confusion matrices for MMI datasets are shown in Figure 8, and their produced metrics are displayed in Figure 11, Figure 14, and Figure 17. The confusion matrices in Figure 9 are displayed for KDEF datasets, and their generated metrics are reported in Figure 12, Figure 15, and Figure 18. Similarly, the confusion matrices for CK+ datasets are shown in Figure 10, and their produced metrics are reported in Figure 13, Figure 16, and Figure 19.

From the precision, it can be seen that the method in the paper does have a significant effect on identifying the probability of being classified as correct, especially in Figures 12 and 13 (KDEF and CK+). From the recall, it can be seen that the proportion of the correctly predicted data in each category is still quite high, especially in Figures 15 and 16 (KDEF and CK+).

F1 takes into account the precision and recall rate. F1 can be seen on the CK+ and KDEF dataset, which show a good performance in Figures 18 and 19. It may be that the MMI dataset contains some occluded images, which leads to its poor performance in precision, recall, and F1. However, compared with other methods in terms of recognition accuracy, our network still has a great advantage as shown in Table 6.

It is clear from Figures 8–19 that the proposed network brings good classification accuracy along with other metrics for KDEF, MMI, and CK+ datasets in almost all the instances. An exhaustive review of these methods is outside the range of our research, and that can be mentioned to [16–18,29–36]. It is clear from Table 8 that Ref. [35] gives good outcomes for the KDEF dataset and AlexNet gives good results for the CK+ data only. However, our recognition rate is completely beyond Broad learning and AlexNet. We can summarize that our network is effective, but it shows the right prediction accuracy in more instances compared to 21 models.

Table 8. Performance comparison of our network with classification accuracy (%) on two datasets viz. CK+ and KDEF.

No.	Method	Ref.	CK+	KDEF
1	AlexNet		97	76
2	HOG-TOP	[15]	65	55
3	SCNN	[28]	61	55
4	MCNN	[28]	85	67
5	SCNN-LBP	[16]	83	70
6	SCNN-gray	[16]	94	78
7	P-VGG16	[16]	91	78
8	WMDC	[16]	97	81
9	WFTS	[29]	91	74
10	ACNN-LBP	[17]	95	66
11	Fusion(ACNN-LBP+GF)	[17]	94	69
12	STF+LSTM.	[18]	82	81
13	Ensemble DCNNs	[30]	67	58
14	DCNN-BC	[31]	73	70
15	IACNN	[32]	95	67
16	2B(N+M)Softmax	[33]	87	81
17	CF+GT	[34]	86	80
18	Broad learning	[35]	81	89
19	Deep-emotion	[36]	94	81
20	VGG19	[37]-1	96	81
21	ResNet150	[37]-2	89	72
22	Our network		99.042	96.53

4.6. Performance Comparison of Execution Time

In comparison, we provide comparative results against 21 state-of-the-art algorithms, for example, HOG-TOP [15], SCNN [28], MCNN [28], SCNN-LBP [16], SCNN-gray [16], P-VGG16 [16], WMDC [16], WFTS [29], ACNN-LBP [17], Fusion(ACNN-LBP+GF) [17], STF+LSTM [18], DCNN-BC [31], IACNN [32], 2B(N+M) Softmax [33], CF+GT [34], Broad learning [35], Deep emotion [36], VGG19 [37]-1, and ResNet150 [37]-2, on two datasets. However, the contrast is limited to average recognition accuracy only. Some methods in Table 9 were executed on videos. Rare works premeditated a fewer number of classes. Table 9 shows the average classification accuracy obtained by the above mentioned 21 methods. It is obvious from Table 9 that our network conquers the twenty-one above models on CK+ and KDEF, and it occurs due to the use of a triple-structure network model. However, the above models are still contrasted according to training and testing time. The training time of a model usually relies on the size of the network, size of the input images, and so on. In the research, the above models were used in the light of their respective specifications. We used tenfold cross validation and 300 epochs when training our network on KDEF and CK+ datasets. The training and testing times required by the above 22 models, including our network on the two databases, are displayed in Table 9. However, testing time for onefold cross validation is noted only in Table 9. Testing time per image (TTPI) is the same for all the images of a database as their size is equal. IACNN takes about 700 min to train KDEF and about 633 min to train CK+, which is quite large. While the proposed method processes an image in 0.278 s, our method is still quite dominant in terms of accuracy.

Table 9. A comparison of the performance of our research in terms of execution time on the two datasets, i.e., CK+ and KDEF.

No.	Method	Ref.	TTPI	Training Time		Testing Time for All the Images	
				CK+	KDEF	CK+	KDEF
1	AlexNet		0.2412	47	51	0.8236	0.8429
2	HOG-TOP	[15]	0.45	92.33	115	1.3264	1.1576
3	SCNN	[28]	0.1812	1.5	2.5	0.8985	0.9001
4	MCNN	[28]	0.1872	2	2.5	0.9127	0.9168
5	SCNN-LBP	[16]	0.1801	16.33	25	0.8991	0.9008
6	SCNN-gray	[16]	0.1801	16.33	25	0.8991	0.9008
7	P-VGG16	[16]	0.4309	158.33	175	1.5238	1.5523
8	WMDC	[16]	0.4699	174.99	200	1.9789	1.9965
9	WFTS	[29]	0.8956	10	16.6	1.9912	1.9989
10	ACNN-LBP	[17]	0.3945	20	30	1.1321	1.1394
11	Fusion(ACNN-LBP+GF)	[17]	0.4612	110	135	1.9984	2.0102
12	STF+LSTM.	[18]	0.4329	215.33	240	2.8628	2.9845
13	Ensemble DCNNs	[30]	0.4917	220	300	2.6296	2.7013
14	DCNN-BC	[31]	0.1725	28	37.33	1.443	1.4523
15	IACNN	[32]	0.3946	633	700	1.9165	1.973
16	2B(N+M)Softmax	[33]	0.2814	85.33	93.33	0.9892	0.9946
17	CF+GT	[34]	0.3218	100	115.33	2.451	2.5503
18	Broad learning	[35]	0.1023	1.5	2	0.3323	0.3812
19	Deep-emotion	[36]	0.2908	50	66.66	0.9165	0.9346
20	VGG19	[37]-1	0.6128	23	26.5	1.9901	2.0981
21	ResNet150	[37]-2	0.7123	67	76.5	2.618	2.7833
22	Our network	model	0.278	482.98	507.2	4.393	4.642

5. Visualization for the Triple-Structure Network Model

To further confirm the effectiveness of the proposed triple-structure network model, a class activation map visualization [67] was performed on each branch of the model. A feature map formed by weighted overlapping of feature atlases can demonstrate the importance of each location to its classification. The first column in Figures 20–22 are the

original images. The second column is the feature channel map (Feature map1) extracted by the global branch network (MobileNet). The third column is the feature channel map (Feature map2) extracted by MobileNet+SENet. The fourth column is the feature channel map (Feature map3) extracted by MobileNet+Div. Finally, the fifth column is the Merged Feature map extracted by our triple-structure network model.

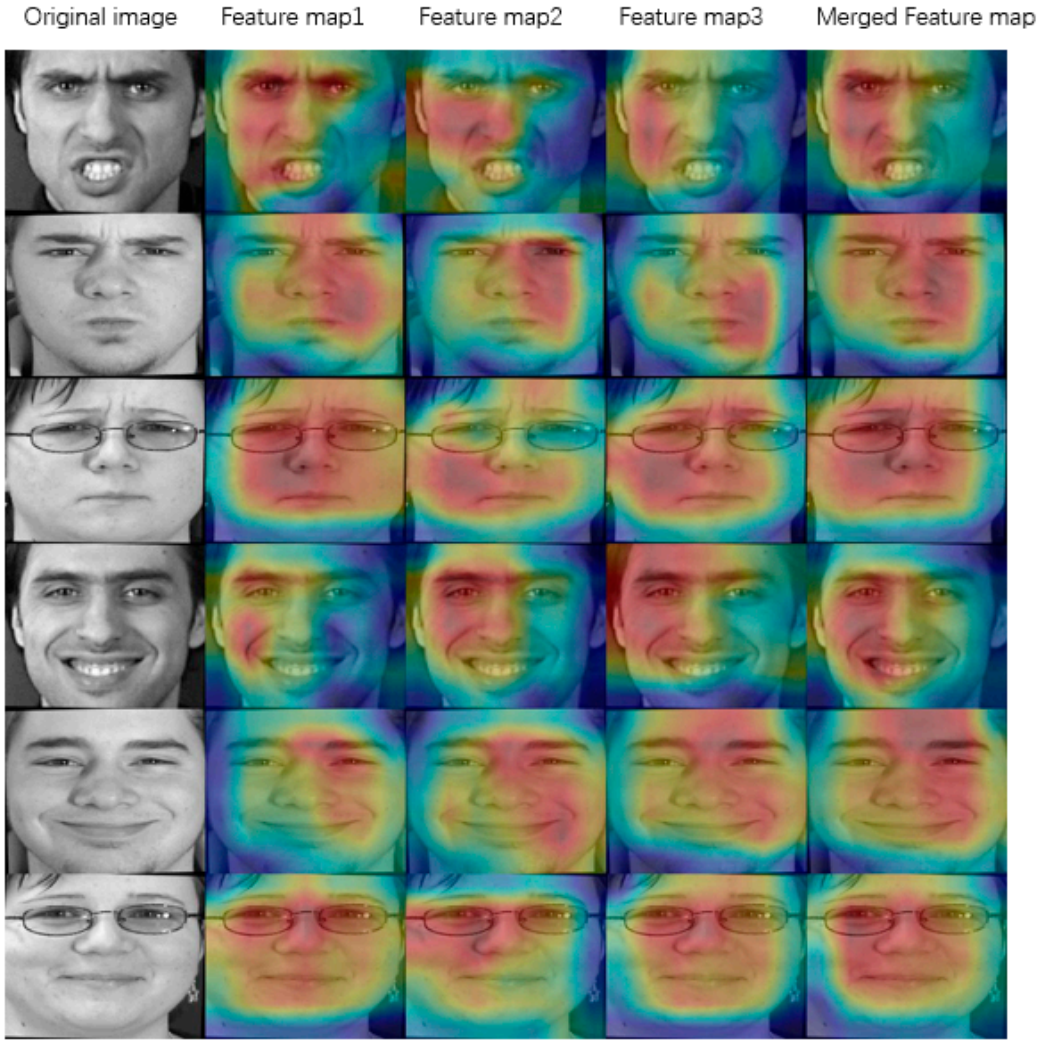

Figure 20. MMI image visualization in each feature channel.

The first three rows of Figure 20 are angry, and the last three rows reflect happiness. The first three rows of Figure 21 are fear, and the last three rows correspond to happiness. The first three rows of Figure 22 are sad, and the last three rows reflect happiness. The second column of Figures 20–22 shows that the global branch network can learn globally about images. In the third column, attention mechanism learning is stressed. In addition, the key learning area of the image is enhanced. The fourth column emphasizes the diverse intra-class, which shows that the learning is performed for multiple regions of the image. The fifth column combines the above advantages for more in-depth learning. The first three rows and the last three rows of Figure 20 are, respectively, angry and happiness, which belongs to an inter-class relationship. It can be observed from Figure 20 that the learning focus on the two types of expressions is different. The key learning area of the angry expression image is around the eyes. The learning area of the expression image is concentrated on the mouth for happiness. This discovery emphasizes the concentration of learning different areas of inter-classes. As for Figure 21, the focus on fear is the eyes and

mouth, and the focus on happiness is around the mouth. As for Figure 22, the focus on sad is the eyebrows and mouth, and the focus on happiness is around the mouth.

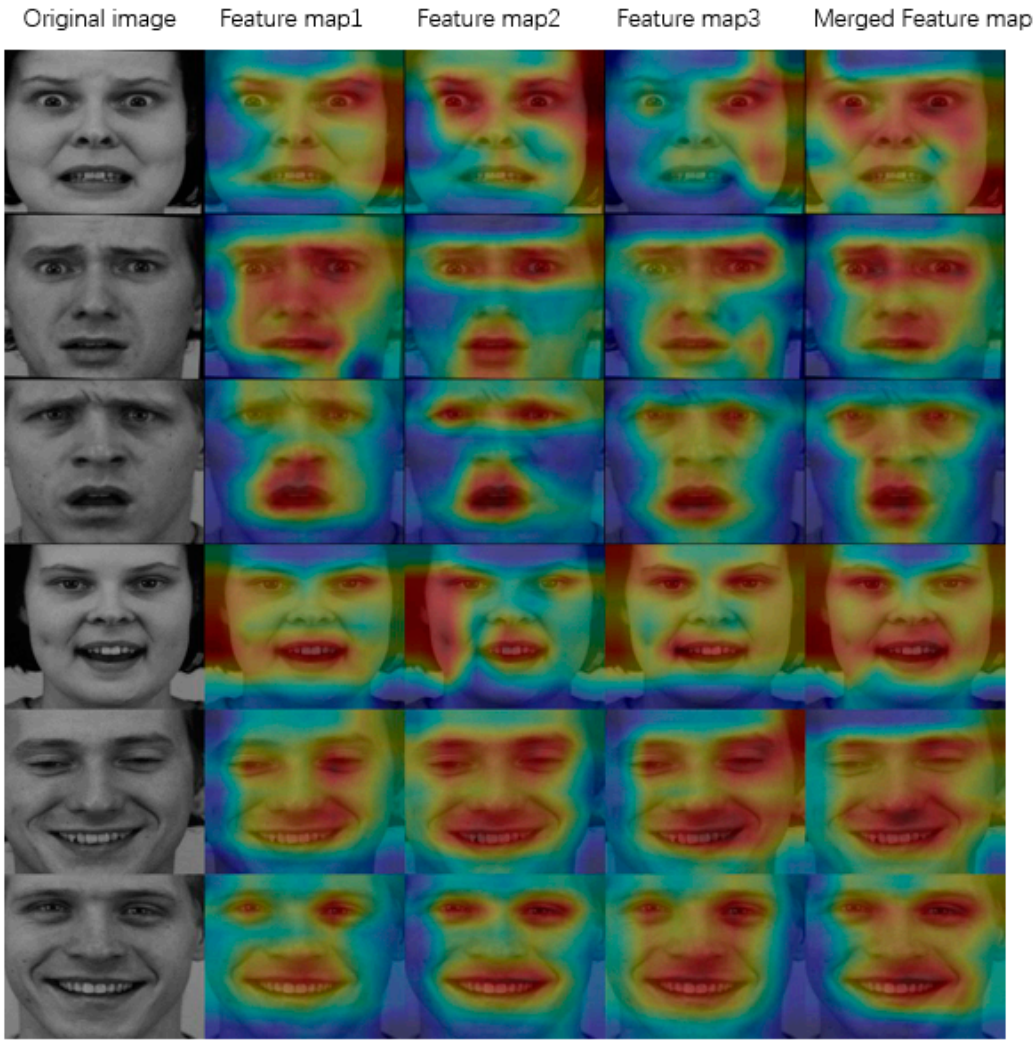

Figure 21. KDEF image visualization in each feature channel.

The expressions for the first three rows in Figures 20–22 are, respectively, angry, fear, and sad. The three are similar in that they focus more on the area around the eyes. Meanwhile, the fear learning has an additional learning towards the area around the mouth and sad has an additional learning towards the area around the eyebrows. The expressions for the last three rows of Figures 20–22 are happiness. This process is focused on the learning of intra-class features. In addition to the mouth, there is also learning around the eyes, which emphasizes the learning of the diversity of intra-class features.

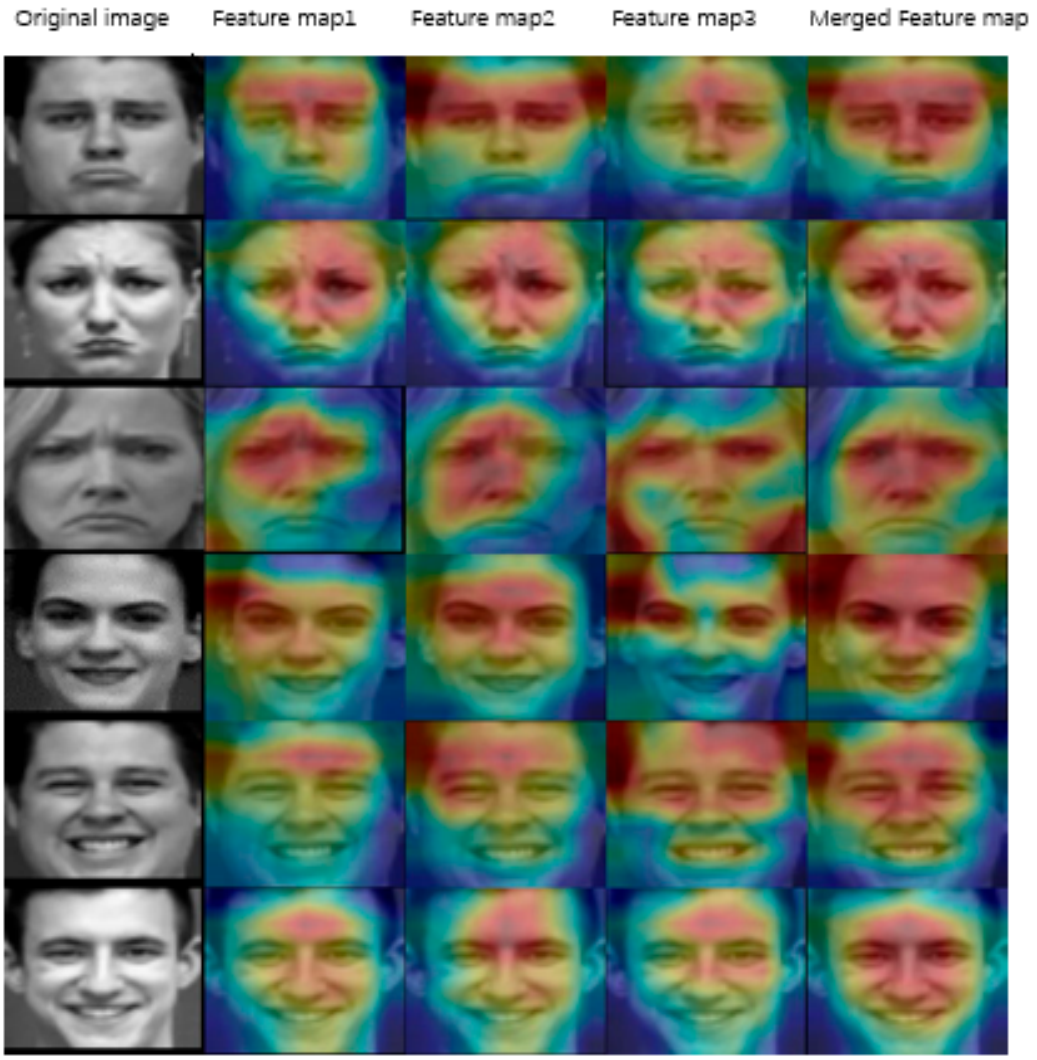

Figure 22. CK+ image visualization in each feature channel.

6. Conclusions

We proposed a symmetric mode to extract the inter-class features and intra-class diversity features, and then put forward a triple-structure network model, which is trained by a new multi-branch loss function. Such a triple-structure network model comprises a global branch network, an attention mechanism branch network, and a diversified feature learning branch network. Our research consists of the below aspects:

(1) We slightly looked back at the previous related works about MobileNet V1 and SENet. Moreover, we highlighted their merits and achievements, which also favors our research idea.

(2) A triple-structure network model was presented. The global branch network focuses on learning global features, the attention branch network concentrates on learning inter-class features, and lastly the diverse feature learning branch network focuses on learning the diversity of intra-class features. Then, the judgment process of the network model comprehensively utilizes global features, inter-class features, and intra-class features. It not only can focus on the overall structure, but also can capture the diverse intra-class and the difference inter-class with a symmetric mode.

(3) Finally, experiments were performed on the KDEF, MMI, and CK+ datasets in which the classification accuracy reached 96.530%, 83.559%, and 99.042%, respectively. Through ablation experiments and visualization, the intrinsic mechanism of our triple-structure network model was proved to be very reasonable. Moreover, experiments

on MMI, KDEF, and CK+ databases demonstrated that our proposed triple-structure network model performs better than dozens of state-of-the-art methods from [10,49], HE+DeepPCA+ELM [11], QIBGSA [12,25], RCFN [38,54], FMN [55], DLP+CNN [56], DeRL [57,58], MF-MLP [60], and HAAR+LDA+IBH-based ELM [61,62].

Based on the experimental results and analyses in Sections 4.3 and 4.4, the proposed network is more competitive than other methods. However, there are still some limitations. For example, our network can only process images, not text and speech. With the development of modern society, the combination of video, voice, and text has been integrated into our world, which shows the limitations of our network.

Although our network applies multiple branches to learn with different focuses, the actual learning is to extract important features of the image to achieve a good recognition rate. However, multi-branch learning can achieve the learning of important features, and it also brings repeated learning and repeated iterations, which results in a waste of computing resources. A key direction we should study is how to propose a network to achieve focused learning of image features, thereby reducing the pressure on computing resources, and enabling better application in practice.

As a next step, we plan to propose a network to achieve focused learning of image features with a reduction on the pressure of computing resources. In the future, it is worthy to research how to construct an expression analysis model with more powerful generalization ability. In addition, we will consider utilizing orthogonal polynomials [68] as a feature extraction tool for facial recognition, which may provide important help in improving the recognition rate.

Author Contributions: Conceptualization, B.H., M.H. and F.R.; methodology, X.W.; software, B.H.; validation, M.H., X.W. and F.R.; formal analysis, F.R.; investigation, B.H.; resources, M.H.; data curation, F.R.; writing—original draft preparation, B.H.; writing—review and editing, M.H.; visualization, X.W.; supervision, M.H.; project administration, F.R.; funding acquisition, X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by National Natural Science Foundation of China under Grant 62176084, and Grant 62176083, and in part by the Fundamental Research Funds for the Central Universities of China under Grant PA2021GDSK0093, PA2022GDSK0068.

Acknowledgments: We acknowledge the use of the equipment provided by the Hefei University of Technology.

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

References

1. Nguyen, T.; Raich, R. Incomplete label multiple instance multiple label learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 1320–1337. [CrossRef]
2. Yang, J.Q.; Chen, C.H.; Li, J.Y.; Liu, D.; Li, T.; Zhan, Z.H. Compressed-encoding particle swarm optimization with fuzzy learning for large-scale feature selection. *Symmetry* **2022**, *14*, 1142. [CrossRef]
3. Tang, Y.; Pedrycz, W. Oscillation-bound estimation of perturbations under Bandler-Kohout subproduct. *IEEE Trans. Cybern.* **2022**, *52*, 6269–6282. [CrossRef]
4. Tang, Y.; Pedrycz, W.; Ren, F. Granular symmetric implicational method. *IEEE Trans. Emerg. Top. Comput. Intell.* **2022**, *6*, 710–723. [CrossRef]
5. Poux, D.; Allaert, B.; Ihaddadene, N.; Bilasco, I.M.; Djeraba, C.; Bennamoun, M. Dynamic facial expression recognition under partial occlusion with optical flow reconstruction. *IEEE Trans. Image Process.* **2022**, *31*, 446–457. [CrossRef]
6. Tang, Y.; Ren, F.; Pedrycz, W. Fuzzy c-means clustering through SSIM and patch for image segmentation. *Appl. Soft Comput.* **2020**, *87*, 105928. [CrossRef]
7. 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.* **2021**, *12*, 544–550. [CrossRef]
8. Wang, S.J.; Yan, W.J.; Li, X.; Zhao, G.; Zhou, C.-G.; Fu, X.; Yang, M.; Tao, J. Micro-expression recognition using color spaces. *IEEE Trans. Image Process.* **2015**, *24*, 6034–6047. [CrossRef]
9. Kaliouby, R.; Robinson, P. Real-time insurance of complex mental states from facial expressions and head gestures. In *Real-Time Vision for Human-Computer Interaction*; Springer: New York, NY, USA, 2005; pp. 181–200.

10. Rujirakul, K.; So-In, C. Histogram equalized deep PCA with ELM classification for expressive face recognition. In Proceedings of the International Workshop on Advanced Image Technology (IWAIT), Chiang Mai, Thailand; 2018; pp. 1–4.
11. Kumar, Y.; Verma, S.K.; Sharma, S. Quantum-inspired binary gravitational search algorithm to recognize the facial expressions. *Int. J. Mod. Phys. C* **2020**, *31*, 2050138. [CrossRef]
12. Islam, B.; Mahmud, F.; Hossain, A. Facial region segmentation based emotion recognition using extreme learning machine. In Proceedings of the International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE), Gazipur, Bangladesh, 22–24 November 2018; pp. 1–4.
13. Xi, X.; Zhang, Y.; Hua, X.; Miran, S.M.; Zhao, Y.B.; Luo, Z. Facial Expression Distribution Prediction Based on Surface Electromyography. *Expert Syst. Appl.* **2020**, *161*, 113683. [CrossRef]
14. Choudhary, D.; Shukla, J. Feature Extraction and Feature Selection for Emotion Recognition using Facial Expression. In Proceedings of the IEEE International Conference on Multimedia Big Data (BigMM), New Delhi, India, 24–26 September 2020; pp. 125–133.
15. Chen, J.; Chen, Z.; Chi, Z.; Fu, H. Facial expression recognition in video with multiple feature fusion. *IEEE Trans. Affect. Comput.* **2016**, *9*, 38–50. [CrossRef]
16. Zhao, X.; Shi, X.; Zhang, S. Facial expression recognition via deep learning. *IETE Tech. Rev.* **2015**, *32*, 347–355. [CrossRef]
17. Kim, J.H.; Kim, B.G.; Roy, P.P.; Jeong, D.M. Efficient facial expression recognition algorithm based on hierarchical deep neural network structure. *IEEE Access* **2019**, *7*, 41273–41285. [CrossRef]
18. Kim, D.H.; Baddar, W.J.; Jang, J.; Ro, Y.M. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Trans. Affect. Comput.* **2017**, *10*, 223–236. [CrossRef]
19. Fei, Z.; Yang, E.; Li, D.; Butler, S.; Ijomah, W.; Li, X.; Zhou, H. Deep convolution network based emotion analysis towards mental health care. *Neurocomputing* **2020**, *388*, 212–222. [CrossRef]
20. Chengeta, K.; Viriri, S. A review of local, holistic and deep learning approaches in facial expressions recognition. In Proceedings of the Conference on Information Communications Technology and Society (ICTAS), Durban, South Africa, 6–8 March 2019; pp. 1–7.
21. He, J.; Li, S.; Shen, J.; Liu, Y.; Wang, J.; Jin, P. Facial expression recognition based on VGGNet convolutional neural network. In Proceedings of the Chinese Automation Congress, Xi'an, China, 30 November–2 December 2018; pp. 4146–4151.
22. Ming, Z.; Chazalon, J.; Luqman, M.M.; Visani, M.; Burie, J.C. FaceLiveNet: End-to-End networks combining face verification with interactive facial expression-based liveness detection. In Proceedings of the 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 3507–3512.
23. Giannopoulos, P.; Perikos, I.; Hatzilygeroudis, I. Deep learning approaches for facial emotion recognition: A case study on FER-2013. In *Advances in Hybridization of Intelligent Methods. Smart Innovation, Systems and Technologies*; Hatzilygeroudis, I., Palade, V., Eds.; Springer: Cham, Switzerland, 2018; Volume 85.
24. Wu, S.; Wang, B. Facial expression recognition based on computer deep learning algorithm: Taking cognitive acceptance of college students as an example. *J. Ambient Intell. Humaniz. Comput.* **2021**, 1–12. [CrossRef]
25. Ye, Y.; Zhang, X.; Lin, Y.; Wang, H. Facial expression recognition via region-based convolutional fusion network. *J. Vis. Commun. Image Represent.* **2019**, *62*, 1–11. [CrossRef]
26. Singh, S.; Nasoz, F. Facial expression recognition with convolutional neural networks. In Proceedings of the 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 6–8 January 2020; pp. 324–328.
27. Chen, X.; Yang, X.; Wang, M.; Zou, J. Convolution neural network for automatic facial expression recognition. In Proceedings of the International Conference on Applied System Innovation (ICASI), Sapporo, Japan, 13–17 May 2017; pp. 814–817.
28. Alizadeh, S.; Fazel, A. *Convolutional Neural Networks for Facial Expression Recognition*; Technical Report; Stanford University: Stanford, CA, USA, 2017; p. 1704.06756.
29. Hua, W.; Dai, F.; Huang, L.; Xiong, J.; Gui, G. Hero: Human emotions recognition for realizing intelligent internet of things. *IEEE Access* **2019**, *7*, 24321–24332. [CrossRef]
30. Pons, G.; Masip, D. Supervised committee of convolutional neural networks in automated facial expression analysis. *IEEE Trans. Affect. Comput.* **2017**, *9*, 343–350. [CrossRef]
31. Villanueva, M.G.; Zavala, S.R. Deep neural network architecture: Application for facial expression recognition. *IEEE Lat. Am. Trans.* **2020**, *18*, 1311–1319. [CrossRef]
32. Meng, Z.; Liu, P.; Cai, J.; Han, S.; Tong, Y. Identity-aware convolutional neural network for facial 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; IEEE: Piscataway, NJ, USA, 2017; pp. 558–565.
33. Liu, X.; Vijaya, K.B.V.; You, J.; Jia, P. Adaptive deep metric learning for identity-aware facial expression recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 20–29.
34. Benitez-Quiroz, C.F.; Srinivasan, R.; Martinez, A.M. Discriminant functional learning of color features for the recognition of facial action units and their intensities. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2835–2845. [CrossRef] [PubMed]
35. Zhang, T.; Liu, Z.; Wang, X.H.; Xing, X.F.; Chen, C.P.; Chen, E. Facial expression recognition via broad learning system. In Proceedings of the 2018 IEEE International Conference on Systems, Man, And Cybernetics (SMC), Miyazaki, Japan, 7–10 October 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1898–1902.

36. Minaee, S.; Abdolrashidi, A. Deep-emotion: Facial expression recognition using attentional convolutional network. *arXiv* **2019**, arXiv:1902.01019. Available online: http://arxiv.org/abs/1902.01019 (accessed on 26 September 2022). [CrossRef]
37. Orozco, D.; Lee, C.; Arabadzhi, Y.; Gupta, D. *Transfer Learning for Facial Expression Recognition*; Technical Report; Florida State University: Tallahassee, FL, USA, 2018; p. 7.
38. Zhou, J.; Zhang, X.; Lin, Y.; Lin, Y. Facial expression recognition using frequency multiplication network with uniform rectangular features. *J. Vis. Commun. Image Represent.* **2021**, *75*, 103018. [CrossRef]
39. Cotter, S.F. MobiExpressNet: A deep learning network for face expression recognition on smart phones. In Proceedings of the IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 4–6 January 2020; pp. 1–4.
40. Nan, Y.; Ju, J.; Hua, Q.; Zhang, H.; Wang, B. A-Mobilenet: An approach of facial expression recognition. *Alex. Eng. J.* **2022**, *61*, 4435–4444. [CrossRef]
41. Ding, Y.; Tang, Z.; Wang, F. Single-Sample Face Recognition Based on Shared Generative Adversarial Network. *Mathematics* **2022**, *10*, 752. [CrossRef]
42. Abdulhussain, S.H.; Mahmmod, B.M.; AlGhadhban, A.; Flusser, J. Face Recognition Algorithm Based on Fast Computation of Orthogonal Moments. *Mathematics* **2022**, *10*, 2721. [CrossRef]
43. Yang, B.; Wu, J.; Hattori, G. Facial expression recognition with the advent of face masks. In Proceedings of the 19th International Conference on Mobile and Ubiquitous Multimedia, Essen, Germany, 22–25 November 2020; pp. 1–3.
44. Sadik, R.; Anwar, S.; Reza, L. AutismNet: Recognition of autism spectrum disorder from facial expressions using Mobilenet architecture. *Int. J. Adv. Trends Comput. Sci. Eng.* **2021**, *10*, 327–334.
45. Petrosiuk, D.; Arsirii, O.; Babilunha, O.; Nikolenko, A. Deep learning technology of convolutional neural networks for facial expression recognition. *Appl. Asp. Inf. Technol.* **2021**, *4*, 192–201. [CrossRef]
46. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
47. Zhong, Y.; Qiu, S.; Luo, X.; Meng, Z.; Liu, J. Facial expression recognition based on optimized ResNet. In Proceedings of the World Symposium on Artificial Intelligence (WSAI), Guangzhou, China, 27–29 June 2020; pp. 84–91.
48. Zhao, Z.; Li, Q.; Zhang, Z.; Cummins, N.; Wang, H.; Tao, J.; Schuller, B.W. Combining a parallel 2D CNN with a self-attention Dilated Residual Network for CTC-based discrete speech emotion recognition. *Neural Netw.* **2021**, *141*, 52–60. [CrossRef]
49. Chang, D.; Ding, Y.; Xie, J.; Bhunia, A.K.; Li, X.; Ma, Z.; Wu, M.; Guo, J.; Song, Y.Z. The devil is in the channels: Mutual-channel loss for fine-grained image classification. *IEEE Trans. Image Process.* **2020**, *29*, 4683–4695. [CrossRef]
50. Goodfellow, I.; Warde-Farley, D.; Mirza, M.; Courville, A.; Bengio, Y. Maxout networks. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; pp. 1319–1327.
51. Pantic, M.; Valstar, M.; Rademaker, R.; Maat, L. Web-based database for facial expression analysis. In Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 6–8 July 2005; p. 5.
52. Lundqvist, M.G. Facial expressions of emotion (KDEF): Identification under different display-duration conditions. *Behav. Res. Methods* **2008**, *40*, 109–115.
53. 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.
54. Mohan, K.; Seal, A.; Krejcar, O.; Yazidi, A. FER-net: Facial expression recognition using deep neural net. *Neural Comput. Appl.* **2021**, *33*, 9125–9136. [CrossRef]
55. Li, S.; Deng, W. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Trans. Image Process.* **2018**, *28*, 356–370. [CrossRef]
56. Yang, H.; Ciftci, U.; Yin, L. Facial expression recognition by de-expression residue learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 2168–2177.
57. Liang, D.; Liang, H.; Yu, Z.; Zhang, Y. Deep convolutional BiLSTM fusion network for facial expression recognition. *Vis. Comput.* **2020**, *36*, 499–508. [CrossRef]
58. Sureddy, S.; Jacob, J. Multi-features Based Multi-layer Perceptron for Facial Expression Recognition System. In *Second International Conference on Image Processing and Capsule Networks. ICIPCN 2021*; Lecture Notes in Networks and Systems; Chen, J.I.Z., Tavares, J.M.R.S., Iliyasu, A.M., Du, K.L., Eds.; Springer: Cham, Switzerland, 2021; Volume 300.
59. Oday, A.H.; Nur, A.A.; Zaheera, Z.A.; Darwish, S.M. Realistic Smile Expression Recognition Approach Using Ensemble Classifier with Enhanced Bagging. *Comput. Mater. Contin.* **2021**, *70*, 2453–2469.
60. Deeb, H.; Sarangi, A.; Mishra, D.; Sarangi, S.K. Human facial emotion recognition using improved black hole based extreme learning machine. *Multimed. Tools Appl.* **2022**, *81*, 24529–24552. [CrossRef]
61. Dubey, A.K.; Jain, V. Automatic facial recognition using VGG16 based transfer learning model. *J. Inf. Optim. Sci.* **2020**, *41*, 1–8. [CrossRef]
62. Meena, H.K.; Sharma, K.K.; Joshi, S.D. Facial expression recognition using the spectral graph wavelet. *IET Signal Process.* **2019**, *13*, 224–229. [CrossRef]
63. Karlekar, A.; Seal, A.; Krejcar, O.; Gonzalo-Martin, C. Fuzzy k-means using non-linear s-distance. *IEEE Access* **2019**, *7*, 55121–55131. [CrossRef]

64. Sharma, K.K.; Seal, A. Modeling uncertain data using Monte Carlo integration method for clustering. *Expert Syst. Appl.* **2019**, *137*, 100–116. [CrossRef]
65. Tang, Y.; Pan, Z.; Pedrycz, W.; Ren, F.; Song, X. Viewpoint-based kernel fuzzy clustering with weight information granules. *IEEE Trans. Emerg. Top. Comput. Intell.* **2022**. [CrossRef]
66. Jain, S.; Seal, A.; Ojha, A.; Krejcar, O.; Bureš, J.; Tachecí, I.; Yazidi, A. A Detection of abnormality in wireless capsule endoscopy images using fractal features. *Comput. Biol. Med.* **2020**, *127*, 104094. [CrossRef]
67. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
68. Mahmmod, B.M.; Abdulhussain, S.H.; Suk, T.; Hussain, A. Fast Computation of Hahn Polynomials for High Order Moments. *IEEE Access* **2022**, *10*, 48719–48732. [CrossRef]