applied sciences

MDPI

Article

Multi-Channel Expression Recognition Network Based on Channel Weighting

Xiuwen Lu [1,2], Hongying Zhang [1,2,*], Qi Zhang [1,2] and Xue Han [1,2]

[1] School of Information Engineering, Southwest University of Science and Technology, Mianyang 621000, China
[2] Sichuan Provincial Key Laboratory of Robotics for Special Environments, Southwest University of Science and Technology, Mianyang 621000, China
* Correspondence: zhywyd@163.com

Abstract: Accurate expression interpretation occupies a huge proportion of human-to-human communication. The control of expressions can facilitate more convenient communication between people. Expression recognition technology has also been transformed from relatively mature laboratory-controlled research to natural scenes research. In this paper, we design a multi-channel attention network based on channel weighting for expression analysis in natural scenes. The network mainly consists of three parts: Multi-branch expression recognition feature extraction network, which combines residual network ResNet18 and ConvNeXt network ideas to improve feature extraction and uses adaptive feature fusion to build a complete network; Adaptive Channel Weighting, which designs adaptive weights in the auxiliary network for feature extraction, performs channel weighting, and highlights key information areas; and Attention module, which designs and modifies the spatial attention mechanism and increases the proportion of feature information to accelerate the acquisition of important expression feature information areas. The experimental results show that the proposed method achieves better recognition efficiency than existing algorithms on the dataset FER2013 under uncontrolled conditions, reaching 73.81%, and also achieves good recognition accuracy of 89.65% and 85.24% on the Oulu_CASIA and RAF-DB datasets, respectively.

Keywords: facial expression recognition; convolution neural network; deep learning

check for updates

Citation: Lu, X.; Zhang, H.; Zhang, Q.; Han, X. Multi-Channel Expression Recognition Network Based on Channel Weighting. *Appl. Sci.* **2023**, *13*, 1968. https://doi.org/10.3390/app13031968

Academic Editors: Zhihan Lv, Kai Xu and Zhigeng Pan

Received: 27 October 2022
Revised: 28 January 2023
Accepted: 28 January 2023
Published: 2 February 2023

1. Introduction

Facial expressions are an important way for humans to express their emotional states. By analyzing and processing learners' facial expressions, the emotional states of learners can be identified [1]. With the improvement of computer computing power and the advancement of neural networks, machine translation, behavior recognition and other technologies, artificial intelligence has made positive progress, starting from simple vision and hearing, and has now entered the stage of human–computer interaction [2]. The automatic recognition of human emotional states is an important issue in human–computer interactions [3,4], and the research on emotional computing has also entered a new journey.

Eye-tracking technology is a research aspect in affective computing that has significant implications for fatigue detection in the security domain. Li et al. [5] used wearable eye-tracking technology to assess the impact of mental fatigue on the operator's hazard detection ability and the corresponding visual attention allocation pattern. Deng et al. [6] proposed a system called DriCare, which uses video images to detect the driver's fatigue state as well as yawning, blinking and eye closing duration. Combining eye-tracking technology with facial expression recognition can also make a difference. Zhan et al. [7] built an intelligent Agent-based emotional and cognitive recognition model for distance learners, coupling eye tracking and expression-monitoring iterative recognition, emotion and cognitive recognition processes in order to improve the recognition accuracy of the distance learner's state and the Agent's emotional and cognitive support to the learner. In

addition, AR/VR technology has also made great progress, especially in the animation industry. AR technology can be used for interactive signing sessions, and a group photo with the movie image [8] through AR technology can be used for immersive experience, which can greatly improve the development of the animation industry.

Emotional computing is inseparable from image recognition, and face tracking and facial feature extraction are the focus of image recognition. Li et al. [9] introduced a multi-target face real-time detection and tracking recognition algorithm which included three methods of fast tracking, fast detection and fast recognition. Zheng et al. [10], proposed an efficient deep learning-based face detection and tracking framework for accurate tracking of faces in video sequences; these include SENResNet face detection and Regression Network-based Face Tracking (RNFT) models. In terms of feature extraction, extracting key features that can summarize facial feature changes can improve the recognition ability of the classifier. Therefore, feature extraction is of great significance to facial expression recognition in the entire recognition and classification process [11]. As long as the features are extracted, the sum and labels are input into the classifier, and a trained classifier is obtained, then the remaining data without labels is classified. According to the different features of expressions, facial expression images can be divided into static image features and dynamic sequence image features. Expression recognition based on static images can extract, classify and recognize facial expression feature information, while dynamic sequence facial expression images contain dynamic information of continuous changes in expressions, which can reflect the changing process of facial expressions [12]. The two different entry methods of static and dynamic have different emphases on the extraction of expression feature information. With the gradual application of convolutional neural networks (CNNs) in the field of image recognition, many scholars use CNN in the field of expression recognition, and continuously adjust the network structure to make facial expression extraction more effective [13]. Facial expression recognition technology has also been further developed in terms of flexibility, artificial dependence, and natural environment.

With the development and deepening of facial expression research, expression recognition methods are becoming more and more diverse. For feature extraction and expression recognition research in uncontrolled environments, scholars have proposed many superior network frameworks. Paleari et al. [14] proposed a multimodal emotion recognition framework based on different possible fusion methods. Cheng et al. [15] studied the expression recognition network in the case of partial face occlusion: In this approach, the image Gabor wavelet features are extracted first, and then the Deep Belief Network (DBN) is used for training and recognition. Expression images are pre-processed with partial occlusion (the occlusion range does not exceed 50%). At the same time, the Gabor wavelet is used to extract features in multiple directions and scales to increase the number of training samples and prevent the network from overfitting. Lv et al. [16] considered that different faces and different parts of the face express different degrees of expression, and used some active expression regions to train a multi-channel deep belief network: First, the face is detected through the sliding window, and then the facial features are detected in turn; then the HOG features of each small area [17] are calculated using the deep belief network for deep learningand the trained features are used as detection components for the Gabor feature fusion. Finally, the network is trained by sparse self-encoding, and the recognition result is obtained. In this way, the redundancy between the expression information can be removed without prior alignment or other preprocessing of the face image. On the facial expression recognition data set, it has obtained a superior recognition effect. Chieh et al. [18] designed a multimodal emotion recognition technique that automatically learns the weighted sum of decision parameters for each modality. Guo [19] and others believe that there is a large amount of redundant information in the facial expression sequence. By calculating the change of the key information points of the face during the expression change process, the face image with the peak expression was selected to determine the peak expression in the face position to improve the accuracy of facial expression recognition.

To solve the problem of facial expression feature recognition with uncertainty, Wang et al. [20] proposed a self-cured network (SCN) where the attention mechanism is added to the classification branch in the network structure, and the idea of regularization sorting is added to reduce the importance of uncertain samples. Finally, in order to further improve the network mechanism, a re-labeling module is added to modify the uncertain sample data. Guo et al. [21] proposed an efficient self-cured network (ESCN) model. A feature weighting network is used to receive facial expression picture video frames and generate class recognition capabilities. then, a multi-scale attention mechanism feature fusion network is used to capture and weight facial expression regions, and finally linear aggregation classification is used for re-label correction.

In summary, the current deep facial expression recognition system has the following problems:

1. Issues such as illumination, occlusion, and large face poses cause certain uncertainty in expression labels under uncontrolled conditions, which affects the accuracy of expression recognition;
2. The database of reliable facial expression recognition is small in scale and poor in quality, which affects the recognition performance;
3. Deep learning involves many hyperparameters, building an expression recognition model takes a long time to train, and the model is large.

Aiming at the above problems, this paper designs and constructs an adaptive expression recognition network based on channel weighting, as shown in Figure 1. The main contributions of this paper are as follows:

1. Design a multi-channel feature extraction network MFE, and construct an expression recognition network in terms of scale, information integrity and fusion methods;
2. Design and use a channel-weighted module ACW, and in the auxiliary branch of MFE, assigns weights to small features from small convolutions to help improve recognition accuracy;
3. Design and modify the spatial attention module and integrate the ACW mechanism to obtain an adaptive spatial attention module ASA, which assists in optimizing the overall network structure to improve the final expression recognition accuracy.

2. Materials and Methods

The main frame of the algorithm proposed in this paper is shown in Figure 1. Specifically, this includes: a Feature extraction network, which builds a multi-channel expression recognition network (MFE) and combines residual network and ConvNeXt [22] network ideas to improve feature extraction; an Adaptive channel weighting module (ACW), which designs adaptive weights in the feature extraction auxiliary network and then performs a Hadamard product with the feature map to weight the channels; and a Self-attention module (ASA), which modifies the spatial attention module and integrates the ACW mechanism to capture key information on facial expressions. Details will be explained below.

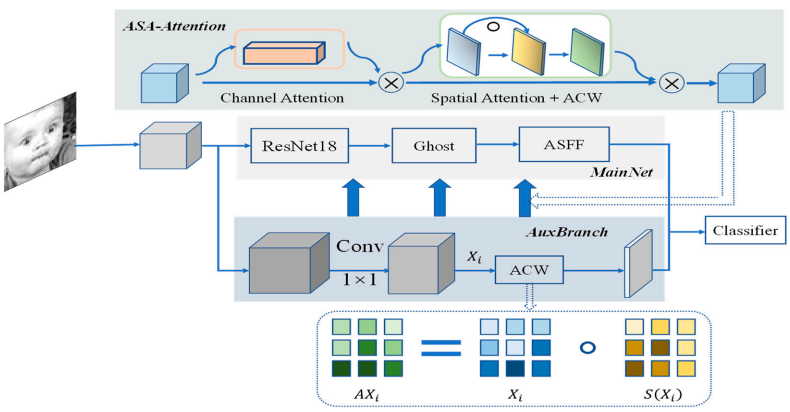

Figure 1. Overall network architecture of MAWNet.

2.1. Multi-Channel Expression Recognition Network

As shown in Figure 1, the overall network consists of the feature extraction backbone network—MainNet—and the feature extraction auxiliary branch—AuxBranch.

2.1.1. MainNet

The backbone of MainNet is built according to ResNet18 [23] and integrates the ConvNext Block. At the same time, considering the large amount of network parameters, the deep structure of ResNet18 is replaced by the Ghost [24] module to complete the main feature extraction network construction. Finally, in order to obtain more complete facial features, feature fusion is performed again using the ASFF [25] structure. The fusion process is shown in Figure 2.

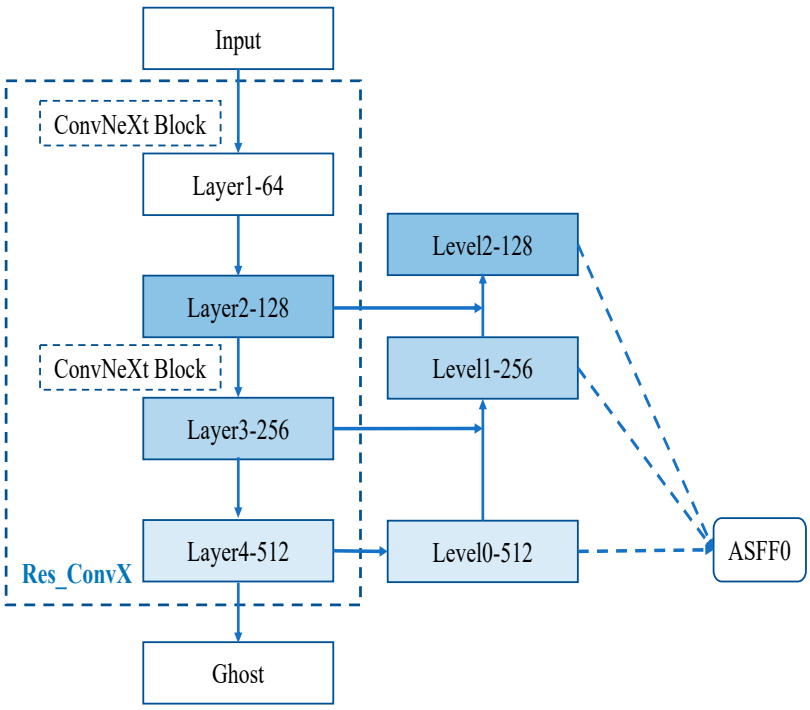

Figure 2. MainNet structure.

The corresponding sizes of the three fused layers are 512 × 5 × 5, 256 × 10 × 10, and 128 × 20 × 20, and the results are pooled by global tie for expression prediction. The formula is as follows:

$$ASFF0 = \alpha_{i,j}^{0} \cdot X_{i,j}^{0\to 0} + \beta_{i,j}^{0} \cdot X_{i,j}^{1\to 0} + \gamma_{i,j}^{0} \cdot X_{i,j}^{2\to 0} \tag{1}$$

*ASFF*0 is used for resizing the feature maps from *Level*1 and *Level*2 to the same size as *Level*0, where $X_{i,j}^{1\to 0}$ represents the feature map after *Level*1 reset, $X_{i,j}^{2\to 0}$ represents the feature map after *Level*2 reset, and $X_{i,j}^{0\to 0}$ represents the feature map from *Level*0.

The values $\alpha_{i,j}^{0}$, $\beta_{i,j}^{0}$ and $\gamma_{i,j}^{0} \in [0,1]$ are the weight parameters of each feature map corresponding to *ASFF*0; the generation of the weight parameter is to reduce the dimensions of the resized *Level1_resized*, *Level2_resized*, and *Level0*, and then splicing. The 1 × 1 convolution is used to calculate $\lambda_{\alpha_{ij}}^{0}$, $\lambda_{\beta_{ij}}^{0}$, $\lambda_{\gamma_{ij}}^{0}$ from *Level1_resized*, *Level2_resized*, *Level*0, respectively. Finally, perform *SoftMax* in the channel dimension and compare it with *Level1_resized*, *Level2_resized*, and *Level0* multiply respectively. The formula is as follows:

$$\alpha_{ij}^{0} = \frac{e^{\lambda_{\alpha_{ij}}^{0}}}{e^{\lambda_{\alpha_{ij}}^{0}} + e^{\lambda_{\beta_{ij}}^{0}} + e^{\lambda_{\gamma_{ij}}^{0}}} \tag{2}$$

Figures 3 and 4 show the feature visualization of whether to add the ASFF module. Here, only 64 channels are randomly selected for display. Comparing Figures 3 and 4, it can be clearly seen that the feature information after adaptive feature fusion retains more complete facial information and discards some invalid channel information, which is more flexible for information fusion and aids to improve the accuracy of facial expression recognition.

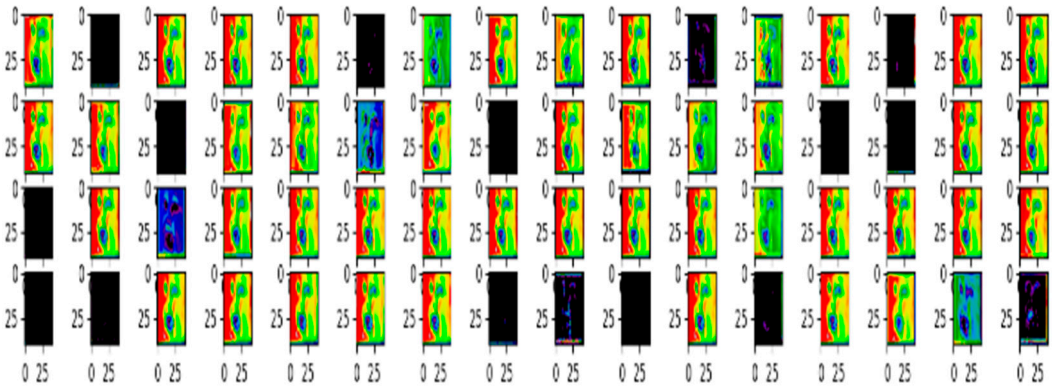

Figure 3. Visualization without ASFF module added.

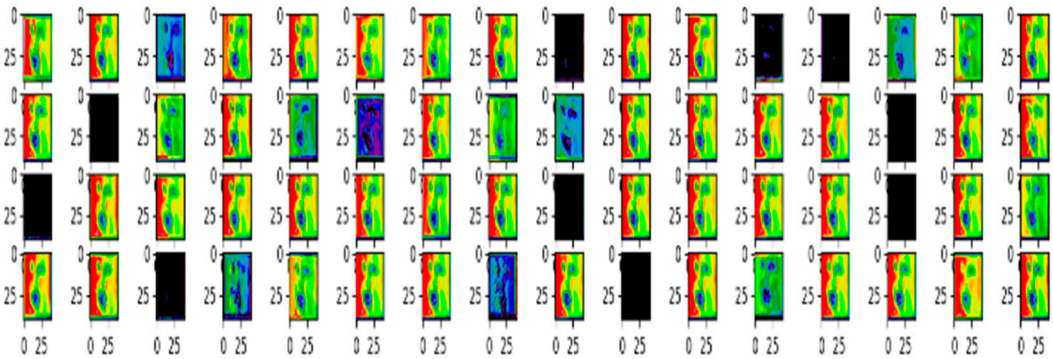

Figure 4. Visualization with ASFF module added.

2.1.2. AuxBranch

The overall structure of AuxBranch is shown in Figure 5 and is mainly composed of a 1×1 convolution. The feature extraction of the secondary branch is performed and the corresponding main network feature information layer is then connected in turn.

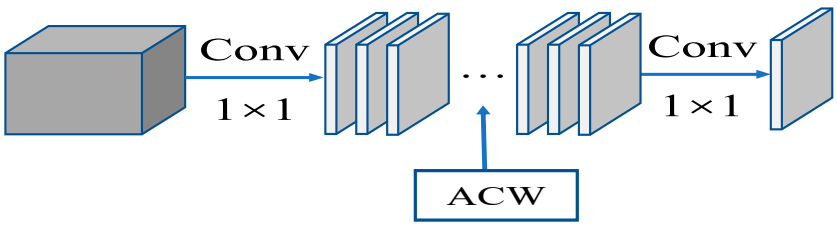

Figure 5. AuxBranch overall structure.

The feature information obtained by AuxBranch is no longer sent to the backbone network but is output as separate feature information, and it is only spliced with the main network feature information at the end. At the same time, the AuxBranch part designs the channel weight ACW, which will be explained in the next chapter.

2.2. Adaptive Channel Weighting—ACW

In the deep neural network, due to the deepening of the network, the feature information will inevitably become more abstract. Therefore, in the multi-channel network design,

this paper selects a 1 × 1 convolution kernel to construct a complete feature extraction network to maximize the possibility of retaining certain input information. In addition, the channel weights are added by design to improve the feature extraction in AuxBranch, as shown in Figure 6.

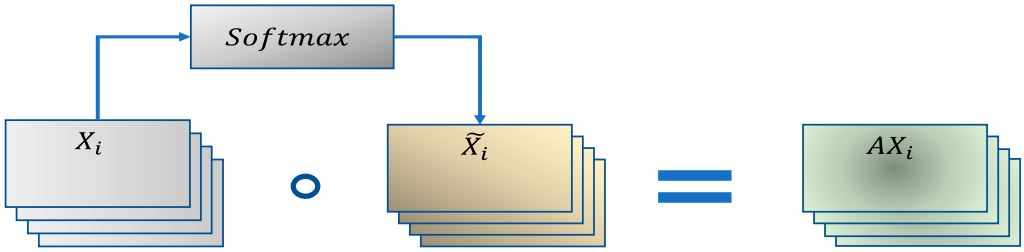

Figure 6. Channel weight diagram.

First, take the output $X_i \in R^{C \times h \times w}$ of the i-th node in the network, where C represents the number of channels of the node, h represents the height, w represents the width, and X_i is composed of C feature maps:

$$X_i = [F_1, F_2, F_3 \ldots F_C] \tag{3}$$

Then perform *Softmax* on X_i to get $\widetilde{X}_i$, which is:

$$\widetilde{X}_i = Softmax(X_i) = \frac{e^{X_i}}{\sum_{C=1}^{C} e^{X_C}} \tag{4}$$

Finally, after obtaining the weight of the channel dimension, use the Hadamard product and the same-dimension feature map for weighting processing:

$$AX_i = X_i \circ \widetilde{X}_i \tag{5}$$

After adaptive channel weighting, the auxiliary recognition network can help the main network to obtain the key information of facial expressions, reduce the weight of irrelevant information, improve the recognition accuracy, and reduce the calculation amount of redundant information. The specific data will be described in the experimental section.

2.3. Attention Module—ASA

The changes in facial expression information are often very subtle and the distinction between some expression categories is not high. At the same time, the facial features between people will also affect the accuracy of expression recognition. Therefore, this paper fuses the attention mechanism ASA to capture key facial information, as shown in Figure 7.

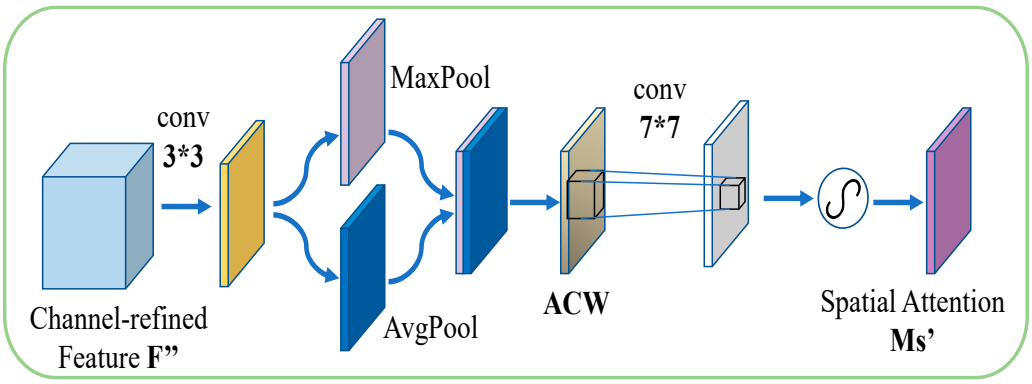

Figure 7. Attention module (ASA) structure diagram.

The ASA module is a fusion modification of the attention mechanism of CBAM [26]. The modification of the ASA module in this paper is only for the spatial attention part, and the channel attention part remains unchanged. Specifically, in the input part, compared with the original module, the ASA module first adds a 3×3 convolutional layer; secondly, after the maximum pooling and average pooling, the channel weighting idea, ACW, is incorporated, and then enters the next convolutional layer.

The 3×3 convolution layer is first added to the original spatial attention module, because the 3×3 convolution has a high degree of non-linearity and can represent more complex functions. After combining with the channel weighting idea, it can have better performance.

3. Experimental Results and Analysis

In this section, the experimental data, the experimental setup, and the validation of the algorithm proposed in this paper are explained, and finally, the overall model is visualized.

3.1. Experimental Data

In order to verify the effectiveness of the network proposed in this paper, this paper selected the public datasets FER2013, Oulu_CASIA and RAF-DB for experimental verification of seven types of basic expressions. Figure 8 shows some pictures from the dataset. The selected dataset covers images of faces in different lighting, skin tones, poses and occlusions.

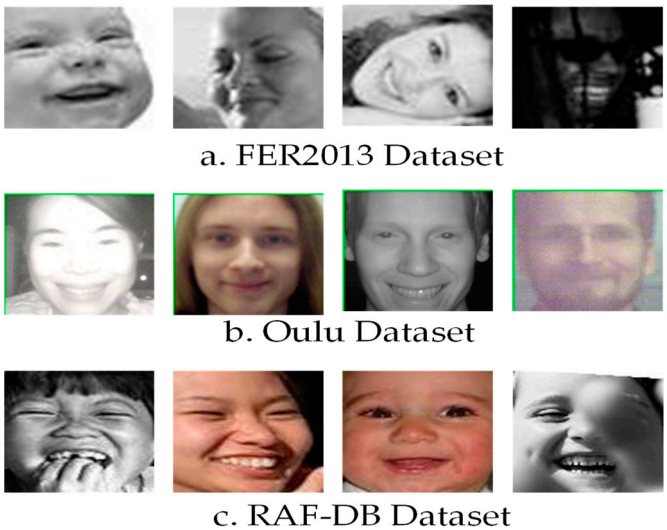

Figure 8. Partial dataset image display.

The FER2013 dataset is a facial expression dataset provided by the Kaggle Facial Expression Recognition Challenge. The dataset has been divided into three parts by the challenge organizers: 28,709 training sets, 3589 public test sets, and 3589 private test sets. The dataset contains facial expressions of different ages and angles, and the resolution is relatively low. Many pictures are also occluded by hands, hair, scarves, etc., which are very challenging and meet the conditions in the real environment.

The Oulu_CASIA dataset contains various expression categories of 80 people divided into two different environmental types, infrared and visible light. The photos were taken under three lighting conditions, ranging from 18 to 30 pictures in each case. In this study, the last six images of each category are taken for training and validation.

The RAF-DB dataset includes 29,672 face pictures in real scenes. Specifically, the RAF-DB dataset has two subsets. The first is a single-label expression dataset, with a total of seven types of expressions; the second is a compound expression dataset, with a total of twelve compound expression categories. This paper studies the problem of single-label facial expression recognition, so the single-label images in the RAF-DB dataset were used to verify the method.

3.2. Experimental Procedure

The experiments in this paper were carried out on NVIDIA GTX 2080Ti CPU. Based on the pytorch deep learning framework, the optimizer adopted SGD, the momentum was set to 0.9, the initial learning rate was set to 0.1, the batch size was set to 16, and the number of training rounds was set to 300 epochs.

In this paper, the data was first preprocessed: the Oulu_CASIA and RAF-DB data sets were detected according to key points, the face part was selected, and the background information was removed; the resolution of the dataset FER2013 itself was 48 × 48, and no other processing was performed. The images in the input model were cropped to a size of 44 × 44, followed by simple data augmentation. Then these were input into the multi-channel feature extraction network for feature extraction, and channel weighting and the attention module were used to help improve the recognition accuracy. Finally, the data was used for expression recognition and classification.

3.3. Ablation Experiment

In order to verify the effectiveness of the method proposed in this paper, each component of the proposed method was experimentally verified, and the FER2013 dataset was selected to evaluate the utility of each module.

The experimental part mainly verified the effectiveness of the proposed multi-channel feature extraction network (MFE), adaptive channel weighting module (ACW), and spatial attention module (ASA). At the same time, in order to prove the applicability of the proposed method to different networks, also considering the network lightweight requirements, this paper extracted Xception [27] to simplify the network: The components of Entry, Middle and Exit remained unchanged, but the operation of each part was not repeated. The following is also called the Xception structure. At the same time, ResNet18 was selected as the Base network in this paper, but the 7 × 7 of the initial convolutional layer was replaced by 3 × 3, which is still called ResNet18 below.

The experimental design is shown in Table 1. Experiment ① is the experimental result in the baseline network ResNet18. Experiment ② is the result obtained in the Xception structure. Experiment ③ is to verify the effectiveness of the attention module ASA proposed in this paper in the Xception structure. Experiment ④ is the effectiveness experiment of adding auxiliary branches (MFE) to the Xception structure. Experiment ⑤ is to verify the effectiveness of the adaptive channel weighting module (ACW) in the Xception structure. Experiment ⑥ is to verify the effectiveness of the ACW method in the Base network. Compared with the original baseline network, the accuracy is improved by 0.46%. Experiment ⑦ is the experimental result after constructing a multi-channel feature extraction network (MFE) on the basis of the baseline network, and the accuracy reaches 73.50%. Experiment ⑧ is to verify the effectiveness of the attention module (ASA) in the baseline network. Compared with the Base network, the accuracy is improved by 0.65%. Experiment ⑨ is the experimental result of the final network structure MAWNet in this paper. It can be seen from the experimental data that the accuracy of MAWNet proposed in this paper is improved by about 1% compared with the baseline network, reaching 73.81%.

Table 1. Performance test of each module.

Method	Base(ResNet)	Xception	MFE	ACW	ASA	Acc
①	√	–	–	–	–	72.88%
②	–	√	–	–	–	71.69%
③	–	√	–	–	√	71.78%
④	–	√	√	–	–	72.11%
⑤	–	√	–	√	–	72.25%
⑥	√	–	–	√	–	73.34%
⑦	√	–	√	–	–	73.50%
⑧	√	–	–	–	√	73.53%
⑨	√	–	√	√	√	73.81%

3.4. Visualization

To explore the structural validity of each module, Figure 9 shows a visualization of the attention module.

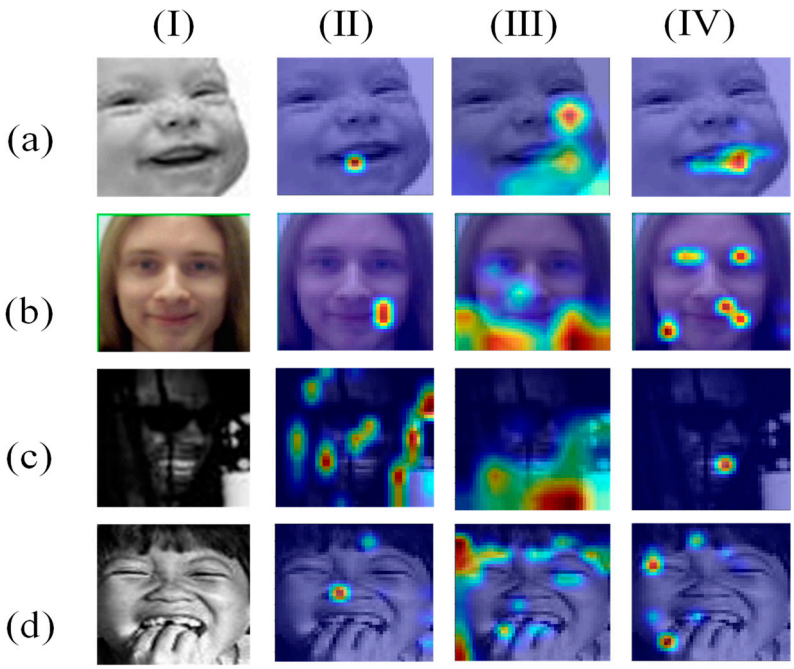

Figure 9. Attention visualization. Where (**a**–**d**) are pictures randomly obtained from the dataset, column (II) is the visualization result of adding ACW, column (III) is the visualization of adding the ASA module, and column (IV) is the visualization result of the network MAWNet proposed in this paper.

First, column (I) randomly selects four pictures a, b, c, and d from the dataset according to lighting, occlusion, and posture. Column (II) corresponds to experiment ⑤, which is the attention visualization result with the addition of ACW mechanism. Compared with the baseline network, the addition of ACW enables the network to pay more attention to expression-related regions; Column (III) corresponds to the above experiment ⑦. In order to add the visualization result of the attention module ASA proposed in this paper, it can be seen from the results of this column that the network can pay more attention to key information areas such as the mouth, eyes and nose; Column (IV) is the attention visualization result of the network MAWNet proposed in this paper. It can be clearly seen that the network better integrates each effective module, can more clearly locate the key areas of expressions, and improve the network recognition accuracy.

4. Discussion

In this section, the overall model is first analyzed and evaluated, followed by experimental validation and comparative analysis of the lightness and classification accuracy of the model, respectively.

4.1. Overall Model Evaluation

In order to visually measure the accuracy of the model and to summarize the prediction results for the classification problem, a confusion matrix was generated and the entire model was evaluated as shown in Figure 10, which briefly analyses the result graphs generated for the FER2013 dataset.

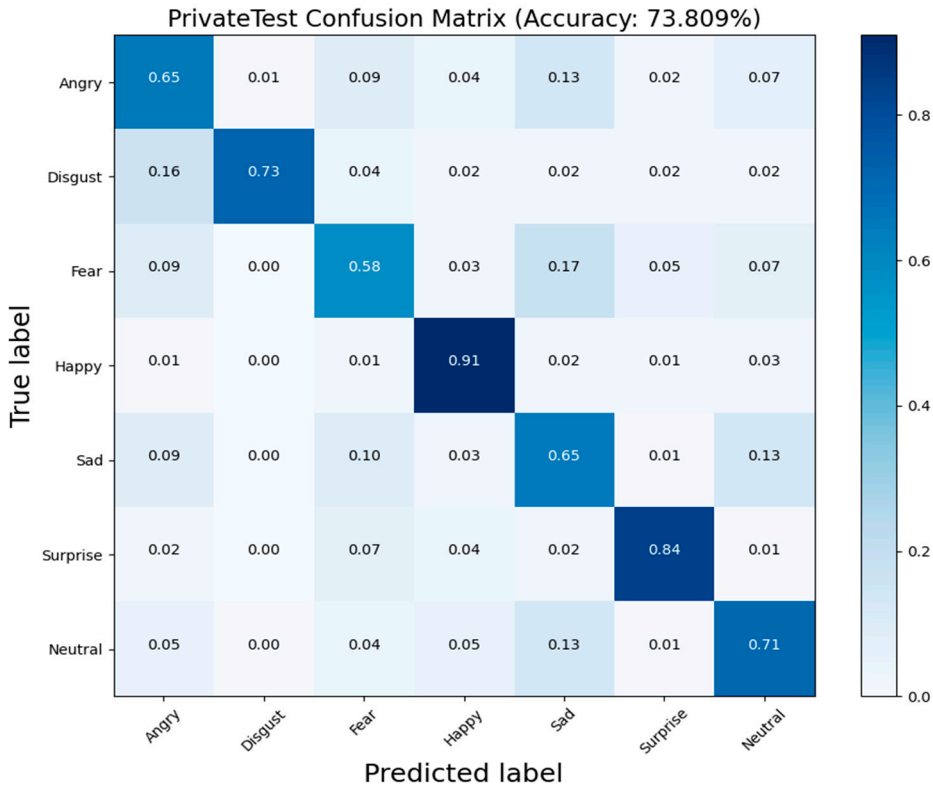

Figure 10. Confusion matrix for expression recognition in the FER2013 dataset.

It can be concluded that the model performs superiorly in the recognition of expressions such as happy, surprised, natural, and disgusted, but slightly less well in expressions such as fear, sad and angry. It is possible to identify the errors made by the classification model and to understand the types of errors made by the model: on the one hand, the uneven distribution of the dataset makes the recognition accuracy of different expression categories different; on the other hand, the great inter-class similarity of expressions such as fear and sadness makes the recognition accuracy of this category low. In addition, the model can also be optimized for the above problems.

4.2. Model Lightweight Experiment

As a common method used in model lightweighting, the advantage of depthwise separable convolution [28] is that the mapping of cross-channel correlation and spatial correlation in the feature map of the convolutional neural network can be completely decoupled. Replacing ordinary convolution with depthwise separable convolution can greatly reduce the number of parameters with comparable accuracy and can be easily defined and modified.

In the model lightweight strategy, this paper also conducted related experiments on the use of depthwise separable convolutions. During the experiment, it was found that, due to the particularity of the expression recognition task, using a too-deep network easily causes overfitting, and the stepwise convolution strategy of depthwise separable convolution will increase the network depth to a certain extent. If overused in this task, network accuracy will be lost. Therefore, the layer4 of ResNet18 was replaced by a deep separable convolution layer in the experiment. This could be achieved with considerable accuracy, greatly reducing the number of parameters.

Table 2 shows the comparison results of the parameters of each method obtained on the FER2013 dataset. It can be seen from Table 2 that, due to the use of depthwise separable convolution in the experiment ② Xception architecture, when the multi-channel feature extraction network is also constructed, the amount of network parameters is 10.48 M, which is 0.69 M less than that of the Base network. Experiment ③ is the beginning of the

construction of the multi-channel feature extraction network in this paper and does not use a lightweight module. Therefore, compared with the Base network, the accuracy is increased by 0.48%, but the number of parameters is increased by 3.33 M. In experiment ④, the network parameters decreased by 4.45 M and the accuracy decreased by only 0.02% after the rational use of the depthwise separable convolution. The result of experiment ⑤ is the multi-channel feature extraction network used in this paper. Compared with the initial structure, the number of parameters is reduced by 4.58 M, and the accuracy reaches 73.50%. Experiment ⑥ shows the complete multi-channel expression recognition network results based on channel weighting in this paper. The parameter quantity and accuracy reaches 73.81%.

Table 2. Parameter comparison.

Experiment	Method	Parameter Quantity(M)	Accuracy
①	Base	11.17	72.88%
②	Xception-MF	10.48	72.11%
③	MF	15.50	73.36%
④	MF_Sep	11.05	73.34%
⑤	MFE(This paper)	10.92	73.50%
⑥	MAWNet(This paper)	11.60	73.81%

4.3. Comparative Experiment

The network framework proposed in this paper has verified its effectiveness on FER2013, Oulu_CASIA, and RAF-DB datasets and shown that it achieves better recognition accuracy. Table 3 shows the implementation of the proposed method on the FER2013 dataset. The experiment compares the recognition effects of some other classic networks and mainstream networks. Among them, the CNN+SVM [29] method uses a convolutional neural network and replaces the SoftMax function at the bottom of the classification network with SVM. The switch is simple, but it is useful for classification tasks. The ARM (ResNet18) [30] method developed a lightweight module to solve the Padding Erosion problem through the auxiliary module method to improve the performance of facial expression recognition. In addition, for the classic network Inception, VGG [31] and the baseline network ResNet used in this paper, the method is also carried out under the same equipment conditions. Finally, compared with FER (2021) [32], using the newer results obtained by various optimization strategies, the method proposed in this paper shows better recognition performance.

Table 3. Accuracy of different methods on the FER2013 dataset.

	FER2013	
Method		Accuracy
Attentional ConvNet [33]		70.02%
CNN+SVM [29]		71.20%
ARM(ResNet18) [30]		71.38%
Inception [31]		71.60%
ResNet [31]		72.40%
VGG [31]		72.70%
FER(2021) [32]		73.28%
MFE (This paper)		73.50%
MAWNet (This paper)		73.81%

Table 4 shows the implementation of the method proposed in this paper on the Oulu_CASIA dataset. The experiment compares some other methods and compares them with some classic depth methods. Deep temporal appearance and geometry network-DTAGN [34] is a two-stream network; a layer of convolutional neural network is used to extract the texture information of the image, and a layer of network is used to focus on the geometric feature changes composed of key points. However, this method is mainly used to recognize

dynamic expressions, so the recognition accuracy is not high. VGG finetune [35] is to fine-tune the classic VGG network, first using large datasets of other tasks for pre-training, and then fine-tuning the expression data. Peak piloted deep network [36] uses samples with greater expression intensity to guide the correct classification of samples with less intensity to improve the recognition rate of expressions with less intensity. FcaeNet2ExpNet [37] uses face recognition datasets to first pre-train the network to retain the basic features of the face, and then uses expression tags to supervise the network to learn expression-related features. Since the identity information of the face is mostly reflected in the facial features, this method has achieved better results.

Table 4. Accuracy of different methods on the Oulu dataset.

Oulu	
Method	**Accuracy**
DTAGN [34]	81.46%
VGG finetune [35]	83.26%
PPDN [36]	84.59%
FaceNet2ExpNet [37]	87.71%
MFE (This paper)	88.34%
MAWNet (This paper)	89.65%

Table 5 shows the comparison of methods on the RAF-DB dataset. The first is the DLP-CNN [38] method, which aims to solve the problem of multiple modal recognition problem. Secondly, ResNet was used as the baseline network and tested under the same equipment conditions. In contrast, the multi-channel recognition network MFE and the complete algorithm MAWNet proposed in this paper achieved better recognition results.

Table 5. Accuracy of different methods on the RAF-DB dataset.

RAF-DB	
Method	**Accuracy**
DLP-CNN [38]	84.13%
Base	84.29%
MFE (This paper)	84.52%
MAWNet (This paper)	85.24%

5. Conclusions

Due to the small feature area of facial expressions, the similarities and differences between feature changes are quite different, so the accuracy of facial expression recognition in natural environment is not high. Therefore, this paper proposes an attentional expression recognition network based on channel weighting. The network constructs a multi-channel structure, aiming to use the auxiliary network to jointly optimize the overall network to improve the convergence speed and recognition accuracy of the network. Firstly, feature extraction and feature fusion are carried out through the subject network. Secondly, the channel weighting module is designed and added, and the auxiliary branch fusion weighting module is used to improve the feature information extraction of the subject network. Finally, the attention module is designed and modified, and the weight of the facial features is increased to achieve the purpose of improving the recognition accuracy and reducing the number of redundant calculations. After the performance verification of each module, and the comparison test and visual analysis of the overall algorithm, the results show that the facial expression recognition method proposed in this paper achieves better recognition accuracy than the existing algorithms. Our method achieved a recognition accuracy of 73.81% on the FER2013 dataset. dataset, and 89.65% and 85.24% on the Oulu_CASIA and RAF-DB datasets, respectively. Due to the completeness of facial feature extraction, the model's processing ability for facial expression recognition in real environments has been enhanced, and its performance has been improved.

However, when recognizing facial expressions in this paper, due to occlusion and insufficient image data of different pose types, the model failed to learn more feature information, which also greatly affected the recognition performance. In future work, we will focus on such problems, and conduct research on the direction of joint optimization to obtain a more superior network model.

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

Funding: This research was funded by the National Natural Science Foundation of China, grant number 61872304.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Li, A.; An, L.; Che, Z. A Facial Expression Recognition Model Based on Texture and Shape Features. *Trait. Du Signal* **2020**, *37*, 627–632. [CrossRef]
2. Gonzalez, J.F.E. Increasing motivation for in-class reading comprehension in a business English course at the University of Costa Rica (UCR). *Res. Pedagog.* **2019**, *9*, 254–265. [CrossRef]
3. Sariyanidi, E.; Gunes, H.; Cavallaro, A. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 1113–1133. [CrossRef]
4. Zeng, Z.; Pantic, M.; Roisman, G.I.; Huang, T.S. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 39–58. [CrossRef]
5. Li, J.; Li, H.; Wang, H.; Umer, W.; Fu, H.; Xing, X. Evaluating the impact of mental fatigue on construction equipment operators' ability to detect hazards using wearable eye-tracking technology. *Autom. Constr.* **2019**, *105*, 102835. [CrossRef]
6. Deng, W.; Wu, R. Real-Time Driver-Drowsiness Detection System Using Facial Features. *IEEE Access* **2019**, *7*, 118727–118738. [CrossRef]
7. Zhan, Z. Intelligent Agent-based Emotional and Cognitive Recognition Model for Distance Learners—Coupling Supported by Eye Tracking and Expression Recognition Technology. *Mod. Distance Educ. Res.* **2013**, 100–105. [CrossRef]
8. Zhang, J.; Li, X.; Chen, K. Analysis of the impact of the animation industry on the development of AR/VR. *Art Sci. Technol.* **2018**, *31*, 104.
9. Li, J.; Wang, Y.; Fang, G.; Zeng, Z. Real-time detection tracking and recognition algorithm based on multi-target faces. *Multimed. Tools Appl.* **2021**, *80*, 17223–17238. [CrossRef]
10. Zheng, G.; Xu, Y. Efficient face detection and tracking in video sequences based on deep learning. *Inf. Sci.* **2021**, *568*, 265–285. [CrossRef]
11. Seng, K.P.; Ang, L.M.; Ooi, C.S. A combined rule-based & machine learning audio-visual emotion recognition approach. *IEEE Trans. Affect. Comput.* **2016**, *9*, 3–13.
12. Bălan, O.; Moise, G.; Petrescu, L.; Moldoveanu, A.; Leordeanu, M.; Moldoveanu, F. Emotion classification based on biophysical signals and machine learning techniques. *Symmetry* **2019**, *12*, 21. [CrossRef]
13. Chen, Y.; Joo, J. Understanding and Mitigating Annotation Bias in Facial Expression Recognition. In Proceedings of the 2021 IEEE International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021.
14. Paleari, M.; Huet, B. Toward Emotion Indexing of Multimedia Excerpts. In Proceedings of the 2008 International Workshop on Content-Based Multimedia Indexing, London, UK, 18–20 June 2008; IEEE: Piscataway, NJ, USA; pp. 425–432.
15. Cheng, Y.; Jiang, B.; Jia, K. A Deep Structure for Facial Expression Recognition under Partial Occlusion. In Proceedings of the Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Kitakyushu, Japan, 27–29 August 2014.
16. Lv, Y.; Feng, Z.; Xu, C. Facial Expression Recognition via Deep Learning. In Proceedings of the 2014 International Conference on Smart Computing (SMARTCOMP), Hong Kong, 3–5 November 2014.
17. 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 (CVPR05), San Diego, CA, USA, 20–25 June 2005; Volume 1, p. 886C893.

18. Huang, K.C.; Lin, H.Y.; Chan, J.C.; Kuo, Y.H. Learning Collaborative Decision-Making Parameters for Multimodal Emotion Recognition. In Proceedings of the 2013 IEEE International Conference on Multimedia and Expo (ICME), San Jose, CA, USA, 15–19 July 2013; IEEE: Piscataway, NJ, USA; pp. 1–6.
19. Yibo, H.; Lingbo, Q.; Lu, W.; Rulong, J. Facial Expression Recognition Based on Adaptive Keyframe Selection. *Inf. Technol.* **2020**, *44*, 19–22.
20. 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 (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 6897–6906.
21. Guo, X.; Ma, N.; Liu, W.; Sun, F.; Zhang, J.; Chen, Y.; Zhang, G. Expression Recognition and Interaction of Pharyngeal Swab Collection Robot. *Comput. Eng. Appl.* **2022**, *58*, 125–135.
22. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A Convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.
23. 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.
24. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. Ghostnet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1580–1589.
25. Liu, S.; Huang, D.; Wang, Y. Learning Spatial Fusion for Single-Shot Object Detection. *arXiv* **2019**, arXiv:1911.09516.
26. 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, 8–14 September 2018; pp. 3–19.
27. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
28. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
29. Tang, Y. Deep learning using linear support vector machines. *arXiv* **2013**, arXiv:1306.0239.
30. Shi, J.; Zhu, S.; Liang, Z. Learning to Amend Facial Expression Representation via De-albino and Affinity. *arXiv* **2021**, arXiv:2103.10189.
31. Pramerdorfer, C.; Kampel, M. Facial expression recognition using convolutional neural networks: State of the art. *arXiv* **2016**, arXiv:1612.02903.
32. Yousif, K.; Chen, Z. Facial Emotion Recognition: State of the Art Performance on FER2013. *arXiv* **2021**, arXiv:2105.03588.
33. Minaee, S.; Minaei, M.; Abdolrashidi, A. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors* **2021**, *21*, 3046. [CrossRef]
34. Jung, H.; Lee, S.; Yim, J.; Park, S.; Kim, J. Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2983–2991. [CrossRef]
35. Ding, H.; Zhou, S.; 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. [CrossRef]
36. Zhao, X.; Liang, X.; Liu, L.; Li, T.; Han, Y.; Vasconcelos, N.; Yan, S. Peak-Piloted Deep Network for Facial Expression Recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland; pp. 425–442.
37. Wang, S.M.; Shuai, H.; Liu, Q.S. Facial expression recognition based on deep facial landmark features. *J. Image Graph.* **2020**, *25*, 0813–0823.
38. Li, S.; Deng, W. Reliable crowdsourcing and deep locality-preserving learining for unconstrained facial expression recognition. *IEEE Trans. Image Process.* **2019**, *28*, 356–370. [CrossRef]