


Article

HSAW: A Half-Face Self-Attention Weighted Approach for Facial Expression Recognition

Shucheng Huang * and Xingpeng Yang 

School of Computer Science, Jiangsu University of Science and Technology, Zhenjiang 212003, China; 211110702108@stu.just.edu.cn

* Correspondence: schuang@just.edu.cn

Abstract: Facial expression recognition plays an increasingly important role in daily life, and it is used in several areas of human–computer interaction, such as robotics, assisted driving, and intelligent tutoring systems. However, the current mainstream methods are based on the whole face, and do not consider the existence of expression asymmetry between the left and right half-face. Hence, the accuracy of facial expression recognition needs to be improved. In this paper, we propose a half-face self-attention weighted approach called HSAW. Using statistical analysis and computer vision techniques, we found that the left half-face contains richer expression features than the right half-face. Specifically, we employed a self-attention mechanism to assign different weights to the left and right halves of the face. These weights are combined with convolutional neural network features for improved facial expression recognition. Furthermore, to attack the presence of uncertain categories in the dataset, we introduce adaptive re-labeling module, which can improve the recognition accuracy. Extensive experiments conducted on the FER2013 and RAF datasets have verified the effectiveness of the proposed method, which utilizes fewer parameters.

Keywords: facial expression recognition; asymmetrical of expression; half-face; self-attention weighted; adaptive re-labeling



Citation: Huang, S.; Yang, X. HSAW: A Half-Face Self-Attention Weighted Approach for Facial Expression Recognition. *Appl. Sci.* **2024**, *14*, 5782. <https://doi.org/10.3390/app14135782>

Academic Editor: Douglas O'Shaughnessy

Received: 8 May 2024

Revised: 25 June 2024

Accepted: 30 June 2024

Published: 2 July 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Facial expressions are a fundamental aspect of human communication, offering critical insights into an individual's emotional state without needing verbal interaction. They are pivotal in social interactions, affecting perceptions, decisions, and behaviors of interpersonal relationships. Therefore, the automatic recognition of facial expressions holds significant promise for enhancing human–computer interaction, contributing to advancements in psychological research, and supporting various applications in the security and entertainment industries. However, accurately recognizing and interpreting facial expressions through computational methods poses a considerable challenge, primarily due to facial movements' complex and dynamic nature.

One of the difficulties in facial expression recognition is the asymmetry of expressions. The reason for this is that the face is the most complex signaling system in the human body and a highly differentiated part of the body [1]. During conversations, Dopson et al. [2] systematically and subtly uncovered that the regions of the cerebral cortex responsible for emotional expressions demonstrate variations between the left and right halves of the face. Such asymmetries often manifested on half or part of the face, introduce discrepancies in expression features, leading to inaccuracies in automated recognition systems. This phenomenon reflects the diverse control mechanisms governing facial muscles and highlights the complexity underlying human emotional expression.

This asymmetry poses a unique challenge to facial expression recognition techniques, as it requires distinguishing and accurately interpreting the features of expression asymmetry. However, asymmetric features may interfere with detection based on the whole

face, and the accuracy of facial expression recognition still needs to be further improved. Recent efforts have focused on leveraging local facial expression features to enhance facial expression recognition accuracy. Liu et al. [3] employed a trained multi-channel convolutional neural network to extract features from localized regions such as the eyebrows, eyes and mouth. Subsequently, these local features are fed into a unified multi-scale fusion network to derive the results used for expression classification. Wang et al. [4] introduced a regional attention network model to address the challenges associated with head deviation and facial occlusion. In their approach, localized regions in a given image are initially identified based on predefined criteria. These regional areas and the original image are then processed through a backbone network to extract features indicative of facial expressions. Ultimately, a classifier leverages the global features, features of the local regions, and their respective weights from the original image to determine the expression category. Although the above methods use local features to improve expression recognition accuracy, the different effects of the left and right half-face on expression classification still need to be further distinguished.

To address the issue aforementioned, in this paper, we propose an expression recognition approach based on a self-attention weighted half-face. In this approach, facial region is divided into the left and right half-faces using a segmentation algorithm. Features are then extracted from each half-face separately using a neural network. Based on the observed asymmetry, appropriate weights are assigned to the left and right halves using a self-attention weighted module. This ensures that the expression features from each half-face uniquely influence the final decision. Consequently, the impact of information asymmetry on the results is minimized. In addition, due to the presence of uncertain classification labels in the dataset after the self-attentive weighting module, this paper introduces an adaptive re-labeling module, which aids in prediction through pseudo-labeling and improves the algorithm's accuracy. Therefore, the main contribution of this paper can be summarized as follows:

- We explore facial expression asymmetry using statistical analysis and computer vision techniques, revealing that the left half-face possesses richer and more recognizable features than the right.
- To address the asymmetry of expressions, we propose a facial expression recognition algorithm based on a self-attention weighted half-face.
- Extensive experiments on the FER2013 and RAF datasets validate the efficacy of the proposed method.

2. Related Works

Traditional facial feature extraction algorithms can be separated into two categories: (1) geometric-based methods, such as Active Appearance Models (AAM) [5], and (2) appearance-based methods, such as LBP [6] and Gabor Wavelet Representation. After the feature description, the features are fed into a classifier, such as SVM [7] and K-nearest Neighbors (KNN) [8], for recognizing different facial expressions. Therefore, the classifier's performance depends mainly on the quality of the extracted features. In [9], various feature extraction techniques combined with different classification algorithms were presented to find the best combination that can be used for emotion intensity recognition. The results with LBP features are better than those using HOG and Gabor features. AAM tracks faces and extracts facial features, and then uses support vector machines to classify facial expressions. In the dataset CK+ [10], the accuracy of this architecture has reached over 65%, with the best recognition accuracy for happy emotions being 100%.

Deep neural networks have become popular in recognizing facial expressions and other computer vision tasks in recent years. VGG Net [11] used very small convolution filters (3×3) to increase the architecture depth, where the small-size filter can make the decision function more discriminative and decrease the number of parameters. The network often stacked several convolutional layers and then followed one pooling layer. The architecture can significantly improve the prior-art configurations when there are

16 or 19 weight layers in the network. GoogLeNet [12] is a 22-layer deep network. The width and depth of the network are increased compared with previous networks. The main structure of the network is the “Inception” layers, which contain several parallel convolution branches. “Inception” layers have different sizes of convolution filters, and the input images can convolve to varying scales of feature maps. In ResNet [13], skip connections are added between the network’s input and output layers. This structure not only increases the training speed and improves the training effect of the model, but also avoids gradient disappearance and network degradation.

Most works were inspired by the above-mentioned deep network architectures for the facial expression recognition task. Sun et al. [14] proposed a facial expression recognition network with visual attention. In this network, the deep convolution features are extracted from the face and, thus, the regions of interest are detected and used to classify expressions. In [15], a GAN-based face frontalization method was presented. The generator formalizes the input face images, and the identity and expression characteristics are preserved at the same time. Then, the discriminator distinguishes between the real images and the generated face images. Liu et al. [3] further considered local facial regions and used trained multi-channel convolutional neural networks to extract local features for eyebrows, eyes, and mouth, thereby improving facial expression recognition accuracy.

However, the current mainstream methods for facial expression recognition are based on the entire face. On the one hand, this method ignores the asymmetry of facial expressions. On the other hand, half-face-based recognition methods can reduce the number of model parameters, making recognition more efficient. Therefore, this paper first verifies the asymmetry of facial expressions from statistical and computer vision perspectives and then proposes an expression recognition algorithm based on a half-face weighted self-attention mechanism.

3. Analysis of the Asymmetry of Facial Expressions

This section analyzes and validates the asymmetry of facial expressions through both statistical analysis and computer vision. We propose a novel approach based on this asymmetry by assigning weights to the left and right half-faces.

3.1. Statistical Analysis

In this subsection, we used statistical methods to analyze the time of human responses during recognition of half-face expressions, ratings of the intensity of half-face expressions, and classification accuracy. This subsection describes the test subjects involved in identification, the stimulus conditions, and the experimental procedure.

3.1.1. Test Subjects

We recruited 18 undergraduate students (8 females) through internet adverts for this experiment (mean age 22.33 years, SD = 1.609 years). All subjects were right-handed, had normal or corrected vision, and had no self-reported history of neurological disease. This paper was approved by the local ethics committee, and conducted according to the principles expressed in the Declaration of Helsinki. All subjects gave written informed consent and were paid for participation.

3.1.2. Stimulus Conditions

To avoid the influence of gender factors on the experiment, half of the male and female expressions were selected as stimulation materials. Additionally, after weighing the emotional sample distribution, resolution, and sample form of different databases, seven types of emotional macro-expression apex images were extracted from the CAS(ME)³ database, with 10 images selected for each type of emotion. Seven of these emotions include happiness, fear, anger, disgust, sadness, surprise, and others. We cropped the front face through OpenFace, then covered the left half-face or right half-face with a white background image for each expression, resulting in 140 emotional stimulation images of

the left and right half-face. They were written into the Psychopy-2023.1.1 program, and presented in the center of the computer screen. Figure 1 illustrates the apparatus and lab environment.

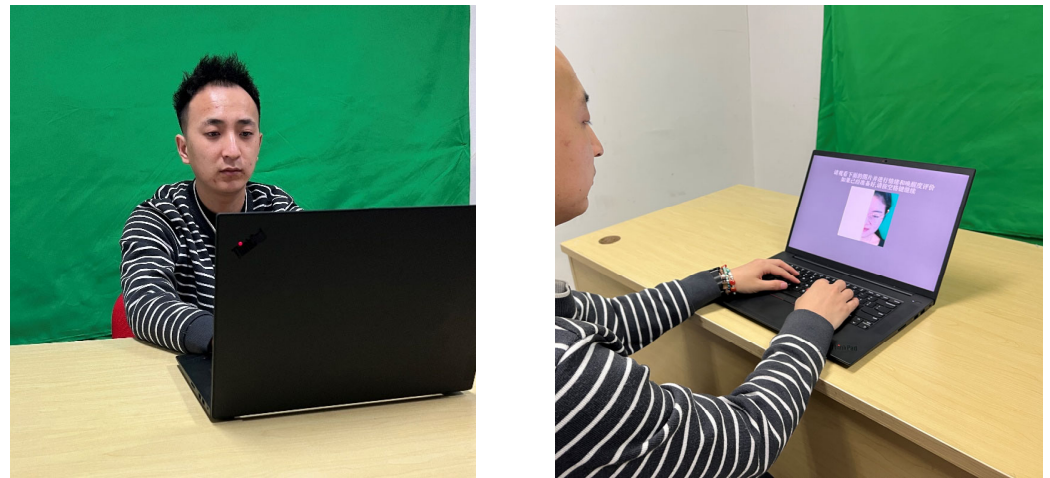


Figure 1. Experimental environment: procedures and laboratory.

3.1.3. Experimental Procedure

Subjects were required to view facial image stimuli and respond to related questions. To exclude practice effects, the order of all trials was counterbalanced across subjects. All subjects used the same program.

In this experiment, subjects were asked to view emotional stimuli displayed on computer screens and answer two questions about emotion types and intensity based on the presented images. Subjects can adjust the distance between themselves and the computer screen while watching the instructions. Still, after entering the formal experimental program, subjects were seated at a distance of 60 cm from the screen with their heads lined up to the center monitor.

Figure 2 presents the experimental procedure in detail. After the formal experiment began, a white “+” fixation point with a duration of 250 ms appeared in the center of the screen, followed by an emotional stimulus image. When the subjects remembered the image, they pressed the “space bar” to answer the question. Each image requires two questions to be answered, and each question is answered using the number keys. The entire experiment requires subjects to complete the judgment of 140 images.

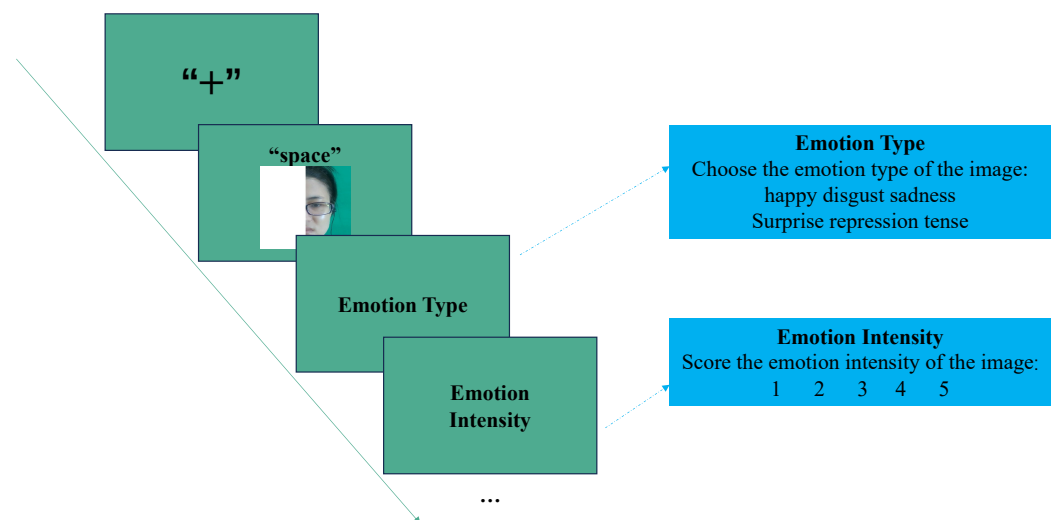


Figure 2. Psychopy program: expression and question presentation sequence. 1–5 represents emotional intensity.

3.1.4. Results of the Analysis

The results of the statistical analyses of human reaction time (RT) during recognition of half-face expressions, ratings of the intensity of half-face expressions, and the accuracy of half-face expression categorization are as follows.

An independent-sample *t*-test revealed no significant differences in RT between the left half-face ($M = 1.461$, $SD = 1.292$) and right half-face ($M = 1.518$, $SD = 1.607$), where $t_{(902.606)} = -0.619$, $p = 0.536 > 0.05$. Furthermore, we classified the emotions correctly recognized by the subjects into positive and negative emotions, with positive emotions only containing happiness and negative emotions including fear, nausea, sadness, and anger. The purpose of doing this is to explore whether positive and negative emotions will have a significant impact on the reaction time of subjects. An independent-sample *t*-test revealed significant differences in RT between the positive emotions ($M = 0.766$, $SD = 0.563$) and the negative emotions ($M = 1.863$, $SD = 1.516$), $t_{(521)} = -11.686$, $p < 0.001$. These results indicate that the RT of positive emotions is significantly less than the RT of negative emotions.

An independent-sample *t*-test for facial emotional intensity revealed a significant difference between the left half-face ($M = 3.25$, $SD = 1.279$) and right half-face ($M = 2.91$, $SD = 1.256$), $t_{(2517.154)} = 6.773$, $p < 0.001$. The above results prove that the emotional intensity of the left half-face is significantly greater than that of the right half-face.

A Chi-square test revealed significant differences in accuracy of discrimination of all emotion types between the left half-face and right half-face, Pearson $\chi^2(1) = 10.552$, $p = 0.001$. The result revealed that the accuracy of the left half-face is significantly higher than that of the right half-face. In addition, we also divided seven emotion types into positive and negative emotions and conducted a Chi-square test on each of them. The results indicate that the accuracy of discrimination positive emotions is still significantly higher on the left face than on the right face, Pearson $\chi^2(1) = 13.425$, $p < 0.001$. However, the accuracy of identifying negative emotions has no significant differences on the left half-face and right half-face, Pearson $\chi^2(1) = 3.318$, $p = 0.069 > 0.05$.

The present Section 3.1 focuses on whether a person observing only the left half of the face and the right half-face affects facial expression recognition. The results showed that the emotional intensity of the left half-face was significantly higher than that of the right half-face, which demonstrated the existence of half-face asymmetry in facial expressions.

3.2. Computer Vision Techniques

Neural networks are one of the core technologies in computer vision. In this subsection, we use two neural network models, AlexNet and ResNet, to analyze the difference between right and left half-face expression recognition.

AlexNet, introduced by Krizhevsky et al. [13], marked a significant breakthrough in the field of deep learning by winning the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012. Its architecture, characterized by deep convolutional layers followed by fully connected layers, was specifically designed to capture and process the intricate patterns and features present in high-resolution images. Liu et al. [13,16–18] found that AlexNet is similar to the primate visual pathway. ResNet, introduced by He et al. [19], further revolutionized the deep learning landscape with its innovative approach to addressing the vanishing gradient problem encountered in very deep networks. Through the introduction of residual connections, ResNet allows for the training of significantly deeper networks by enabling the seamless flow of gradients. We therefore used them to train the left and right half-face expressions separately.

During training, we employed two data augmentation techniques to expand the dataset and improve the robustness of our model. Random rotation: Each image is randomly rotated within a range of -20° to $+20^\circ$. This helps the model become invariant to slight changes in orientation. Random erasing: We applied random erasing to each image, where a randomly selected rectangular region is erased (set to zero). This encourages the model to focus on the most informative parts of the face and improves generalization. In

the classification task, the images are categorized into seven different classes and, hence, the output dimension of the fully connected layer is adjusted to seven. A few original samples from each dataset are shown in Figure 3, and a summary of the facial expression datasets is shown in Table 1. Optimization is performed using Stochastic Gradient Descent (SGD) with the learning rate set at 0.01, momentum decay at 0.9, and the weight decay parameter fixed at 0.0005. The evaluation metric for classification task is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where TP: True Positive, TN: True Negative, FP: False Positive, and FN: False Negative. Experiments from different networks are also conducted using the same setting environment, such as hyperparameters, preprocessing, and augmentation, as well as evaluation metrics.

The CrossEntropyLoss function is utilized as the loss mechanism. The results of their classification are shown in Table 2:

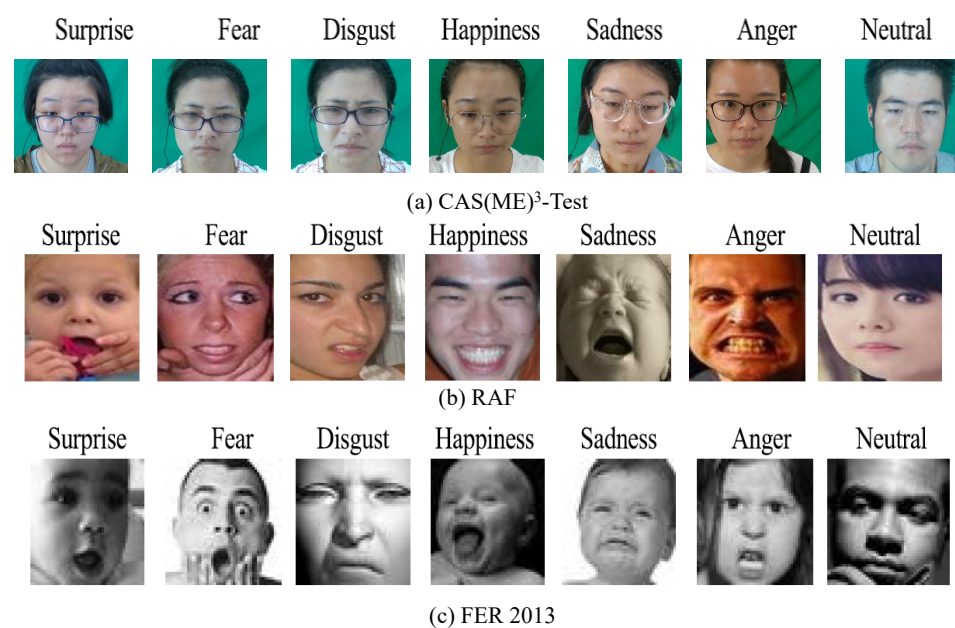


Figure 3. Sample images of facial expression datasets.

Table 1. Summary of facial expression datasets.

Datasets	Samples	Class	Resize	Validation Methods
CAS(ME) ³ -Test	140	7	100 × 100	10-fold cross-validation
FER-2013	35,887	7	48 × 48	10-fold cross-validation
RAF	15,539	7	100 × 100	10-fold cross-validation

Table 2. Comparison of left and right half-face expression recognition accuracy (ACC (%)).

Method	Left ACC	Right ACC
AlexNet	77.6 ± 0.56	73.7 ± 1.32
ResNet	84.7 ± 0.43	79.3 ± 0.76

Table 2 shows that both AlexNet and ResNet networks exhibit higher accuracy in classifying the left half of facial expressions compared to the right half of facial expressions.

To enhance the credibility of our conclusions, we opted for the widely recognized FER2013 dataset as an alternative stimulus material for validation. We conducted five

independent training runs, each using different random initial conditions. This can more accurately reflect the performance and stability of the model, as shown in Table 3.

Table 3. Comparison of facial expression classification expression recognition accuracy (ACC (%)) utilizing the FER2013 dataset within the ResNet.

Datasets	ALL_ACC	Left_ACC	Right_ACC
FER 2013	67.2 ± 0.63	65.7 ± 0.93	60.4 ± 1.03

The ensuing results continued to affirm our initial findings. The detailed outcomes are depicted in Figure 4.

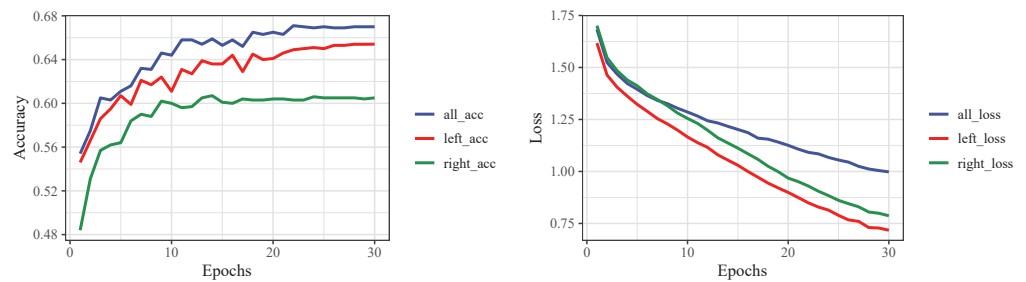


Figure 4. Comparison of facial expression classification outcomes utilizing the FER2013 dataset within the ResNet.

3.3. Section Summary

In this subsection, we find that the left half-face produces more effects than the right half-face in the facial expression recognition process, statistically for human recognition of the left and right half-face expressions and the classification of the left and right half-face expressions by computer vision methods. This observation confirms the existence of asymmetry in facial expressions.

4. Proposed Method

In this paper, we propose an expression recognition approach based on a self-attention weighted half-face; the specific structure is shown in Figure 5. This approach comprises five distinct modules: image segmentation, a feature extraction network, self-attention weighted, bilinear prediction fusion, and adaptive re-labeling. Notably, image segmentation, weighted self-attention, and bilinear prediction fusion are integrated into a self-attention auxiliary module designed to enhance the efficacy of expression recognition. Furthermore, the adaptive re-labeling module is critical in accurately identifying expressions with uncertain labels within the dataset.

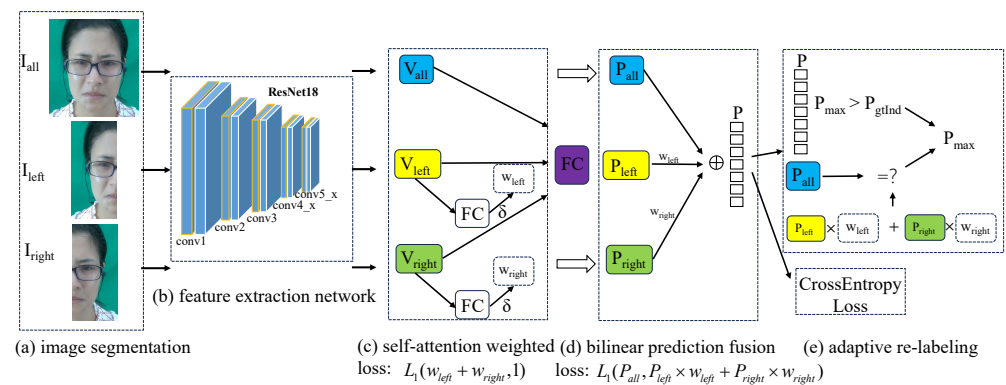


Figure 5. HSAW: half-face self-attention weighted approach: (a) image segmentation, (b) feature extraction network, (c) self-attention weighted, (d) bilinear prediction fusion, and (e) adaptive re-labeling.

4.1. Image Segmentation

In the preprocessing stage for facial expression images within the dataset, the Dlib library is employed to isolate the facial region and identify 68 facial landmarks. Five facial key points are selected for their significance: the tip of the nose, the centers of both eyes, and the corners of the mouth. Utilizing the delineated facial frame denoted as I_{all} , the process involves selecting the key point at the tip of the nose as the pivotal boundary marker. Subsequently, the facial frame is divided into left and right half-face regions, labeled as I_{left} and I_{right} , respectively, through a vertical and symmetrical segmentation.

4.2. Feature Extraction Network

This paper employs a feature extraction network derived from ResNet18 to address the above-mentioned challenges. It confines its selection to the con1 through con5_x layers of ResNet18 for feature extraction, intentionally excluding the global pooling layer to minimize the loss of information. For a specified batch of images, denoted as I_i , this network is utilized to secure the $512 \times 7 \times 7$ feature representation, labeled F_i^5 .

However, the absence of a global pooling layer in the extraction network means that using this feature as input to a fully connected layer for direct classification can result in issues like excessive parameters and overfitting. This paper opts for large convolutional kernel operations instead of global pooling to mimic the effect of dimensionality reduction and prevent overfitting. Firstly, employing large convolutional kernels can substantially augment the model's receptive field, thereby broadening the scope of input features that can be captured. Secondly, the use of large convolutional kernels introduces greater shape biases to the model, which can help in distinguishing between different facial expressions by emphasizing the structural characteristics of the input images [20]. Given that the output dimensions of the con5_x layer are $512 \times 7 \times 7$, employing a convolution kernel larger than 7×7 would result in the kernel sliding only once across the feature map. To circumvent this limitation, sub-pixel convolution is utilized to effectively increase the feature map's spatial dimensions, enhancing its length and width. This technique allows for more detailed processing of the feature map while maintaining the integrity of the extracted features [21]. Specifically, the output feature map from the con5_x layer is initially up-sampled by a factor of 8, yielding a feature map with dimensions of $8 \times 56 \times 56$. Subsequently, a convolution operation utilizing a 16×16 kernel size is executed with a stride of 4, producing a resultant feature map of dimensions $1 \times 11 \times 11$. A reshape operation is conducted to acquire a feature vector V_i^5 with dimensions of 121×1 to finalize the process.

4.3. Self-Attention Weighted

V_{all} , V_{left} , and V_{right} are three feature vectors extracted by the feature extraction network from I_{all} , I_{left} , and I_{right} , respectively. Owing to facial asymmetry, the intensity of emotional expressions varies between the left and right sides of the face which, in turn, influences the final classification outcomes differently. To encapsulate this variance within the model, this paper employs a self-attention mechanism that effectively acknowledges and integrates the distinct contributions of each half-face to emotional expression recognition. The self-attention weight module comprises a fully connected layer (FC) and a sigmoid activation function. This configuration derives expression information from the half-face region's local feature vectors V_{left} and V_{right} and quantifies it as weight values. These weights digitally represent the contribution of the half-face's local features towards the final prediction outcome in the classification process, thereby encapsulating the nuanced impact of each half-face on emotion recognition.

The contribution of each local feature vector V_i , denoted as the weight w_i within the self-attention mechanism module, is calculated using the following approach:

$$w_i = \sigma(W^T V_i), i \in \{left, right\} \quad (2)$$

where $V_i \in R^{(121 \times 1)}$ represents the local feature vector of the half-face and $W \in R^{(121 \times 1)}$ represents the parameters of the fully connected layer. $\sigma(\cdot)$ represents the sigmoid activation function. In contrast, the entire face I_{all} possesses a wealth of global information; thus, its contribution score is assumed to be fixed at 1. To formalize the relationship between the half-face and the entire face, a loss function has been designed, expressed as follows:

$$\mathcal{L}_w = L_1(w_{left} + w_{right}, 1) \quad (3)$$

where $L_1(\cdot)$ represents a smooth ℓ_1 loss function [21]. The value of 1 constrains the weights of the left and right faces, such that they approximate the weight of the complete face in terms of their impact on the final prediction outcome. This loss function is instrumental in guiding the module to allocate weights within defined limits, ensuring that the contributions of the half-faces are balanced and aligned with the comprehensive facial information. This balancing act facilitates a more accurate and holistic interpretation of facial expressions by acknowledging the integral role of both global and local facial features in emotion recognition.

4.4. Bilinear Prediction Fusion

The bilinear prediction fusion module processes the features extracted by the feature extraction network in two significant aspects: making a global prediction, P_{all} , based on the complete facial features, V_{all} . Based on the local features of the left and right facial regions, a local prediction is made using the formula:

$$P_i = W_{fc}^T V_i, i \in \{left, right\} \quad (4)$$

where $W_{fc} \in R^{(121 \times n)}$ represents the parameter matrix of the fully connected layer; n represents the category of expressions. Based on the half-face region for auxiliary prediction P_{aux} , then combining P_{aux} with global prediction P_{all} , the calculation process is as follows:

$$P_{aux} = P_{left} \times w_{left} + P_{right} \times w_{right} \quad (5)$$

$$P = P_{aux} + P_{all} \quad (6)$$

where $P \in R^{(n \times 1)}$ represents the model's ultimate prediction output, wherein the facial expression category with the highest probability is selected as the final prediction outcome.

To augment the feature extraction network's proficiency in gleaning half-face feature information, a smooth L_1 loss function is employed. This mechanism compels the model to strike an optimal balance between learning local and global features. The calculation process of the feature balance loss function is as follows:

$$\mathcal{L}_F = L_1(P_{all}, P_{aux}) \quad (7)$$

Through the backpropagation of the loss function, the model is forced to deepen its preference for local features with higher weights. In addition, to ensure that the features learned by the model are related to expressions, the cross-entropy loss function is constructed directly using the one-hot vector P_{gt} composed of the model's output P and the actual labels.

$$\mathcal{L}_C = - \sum_{k=0}^{n-1} P_{gt}^{(k)} \times \log P^{(k)} \quad (8)$$

The total loss function of the final model is:

$$\mathcal{L} = \alpha \times \mathcal{L}_w + \beta \times \mathcal{L}_F + \gamma \times \mathcal{L}_C \quad (9)$$

where α , β and γ , respectively, represent their corresponding proportions. To enhance the model's feature learning capability, it is essential to meticulously fine-tune the balance between the feature balance loss function and the cross-entropy loss function.

4.5. Adaptive Re-Labeling

Accordingly, this paper draws upon the re-labeling module outlined in [22], leveraging the intrinsic information contained within images to assist the model in accurately classifying ambiguously presented expressions. The re-labeling module employs a dual-dimensional approach for improved recognition accuracy: initially, it evaluates the congruence between the aggregate of global and auxiliary predictions and the original label to identify discrepancies; subsequently, it assesses the consistency between the global prediction and the auxiliary prediction. Only when both criteria are satisfactorily met is the algorithm permitted to assign a pseudo-label to the image. Specifically, if the predicted category in P does not align with the actual label, the adaptive re-labeling module assesses whether the auxiliary prediction (P_{aux}) matches the prediction for the entire face (P_{all}). Should these predictions concur, the expression is assigned the prediction with the highest probability as a pseudo-label. The module refrains from any intervention in scenarios where predictions do not match. This approach enhances the model's performance on datasets featuring ambiguous labels by elucidating the relationship between auxiliary and comprehensive global predictions.

5. Experiments

In this section, we evaluate and compare our method to baselines on two commonly used expression datasets. In addition, ablation studies are performed on individual model components and compared to state-of-the-art models. Extensive experiments on the FER2013 and RAF datasets validate the efficacy of the proposed method.

5.1. Configurations and Datasets

To ensure the reproducibility of our experiments, we provide detailed information about the hardware and software environment used. Table 4 shows the specific hardware and software configurations.

Table 4. Hardware and software details.

CPU	GPU	Memory	Programming Language	Deep Learning Framework
Intel Core i7-8750H (Santa Clara, CA, USA)	NVIDIA GeForce RTX 1050Ti (Santa Clara, CA, USA)	16GB DDR4	Python 3.9	Pytorch 1.7

The ResNet18 network utilized in this paper was pre-trained on the ImageNet dataset, with the following specific hyperparameters established: the batch size was set to 128; the learning rate was adjusted to 0.001; the coefficients α , β , and γ within the total loss function were set to 3, 1, and 2, respectively; and the model underwent a total of 50 training epochs. Table 5 shows the specific parameters for each model.

Table 5. Machine learning model parameters.

Batch Size	Learning Rate	α	β	γ	Epoch
128	0.001	3	1	2	50

Starting from the 20th epoch, the adaptive re-labeling module was incorporated into the training process, culminating in applying 10-fold cross-validation. Descriptions of the used datasets are listed below.

FER2013: The FER2013 database was originally published in the International Conference on Machine Learning (ICML) in 2013 [23]. This dataset consists of 35,887 pictures of faces with 48×48 pixels in grayscale, all images are labeled as seven facial expressions and distributed as follows: 4953 angry images, 547 disgust images, 5121 fear images, 8989 happy images, 6077 sad images, 4002 surprise images, and 6198 neutral images.

RAF: The Real-world Affective Face (RAF) database emerges as a novel dataset that closely mirrors real-world conditions, proving valuable for Multimodal Facial Expression

Recognition (MFER) methodologies as well [24]. RAF comprises 12,271 training samples and 3068 test samples, all of which were annotated by 315 human coders. To mitigate variability stemming from individual annotator biases, the final annotations were refined using crowdsourcing techniques. Within the RAF dataset, facial images are categorized into seven classes, encompassing six basic expressions—anger, disgust, fear, happiness, sadness, surprise, and neutral.

5.2. Ablation Studies

In the ablation study, we systematically evaluate the performance impact of various components on top of the ResNet18 backbone. The components evaluated encompass a self-attention mechanism designed to augment the model's proficiency in prioritizing varying weights of the half face, along with an adaptive re-labeling strategy intended to refine training labels according to the model's predictions. Specifically, we compare the following configurations: (1) ResNet18 as the baseline, (2) ResNet18 augmented with self-attention weights, (3) ResNet18 enhanced with adaptive re-labeling alone, and (4) ResNet18 combined with both self-attention weights and adaptive re-labeling. This structured approach allows us to dissect the contribution of each component to the overall performance.

As illustrated in Table 6, the incorporation of additional modules markedly enhanced the model's performance in facial expression classification (refer to lines 2 and 4). Conversely, the sole addition of the adaptive re-labeling module yielded little improvement in recognition performance (see line 3), possibly due to the model's inadequacy in adeptly learning expression features without differentiating between the weights of the left and right sides of the face. Thus, introducing pseudo labels to images may detrimentally affect the model's performance. This improvement underscores the effectiveness of integrating a self-attention mechanism and an adaptive re-labeling module, facilitating the model's enhanced capability to learn facial expression features by assigning distinct weights to different half-faces.

Table 6. The impact of model component configuration on the recognition accuracy (ACC (%)) of FER2013 and RAF datasets.

Components	FER2013 ACC	RAF ACC
ResNet18	66.7	85.7
ResNet18 + Self-attention	69.4	90.0
ResNet18 + Adaptive Re-labeling	65.9	85.1
ResNet18 + Self-attention + Adaptive Re-labeling	71.1	91.0

5.3. Compared with Other Methods

To ensure a fair comparison between our method and existing state-of-the-art methods, we reproduced the best-performing method listed on paperswithcode and the methods that have been widely used in recent years.

We conducted five independent experimental runs for each reproduced method, and used the mean performance as the final result. This process ensures the stability and reliability of the experimental outcomes. In addition to recording the accuracy of facial expression classification, we also documented the parameter count of each model. The detailed findings from this comparative analysis are delineated in Tables 7 and 8.

As shown in Table 7, facial expression recognition accuracy on the FER 2013 dataset reached a maximum of 74.3%, but the corresponding parameter count was 140.3 million. In addition, the accuracy of EmoNeXt was 72.1%, with a parameter count of 93.6 million. Compared to these methods, although our method has slightly lower accuracy, its parameter count is approximately one-quarter and one-third of theirs, respectively. Similarly, in Table 8, our method differed from the optimal method by only 1% in accuracy, but the number of parameters in our model was significantly lower. These results indicate that our proposed expression recognition algorithm based on the half-face weighted self-attention mechanism is effective.

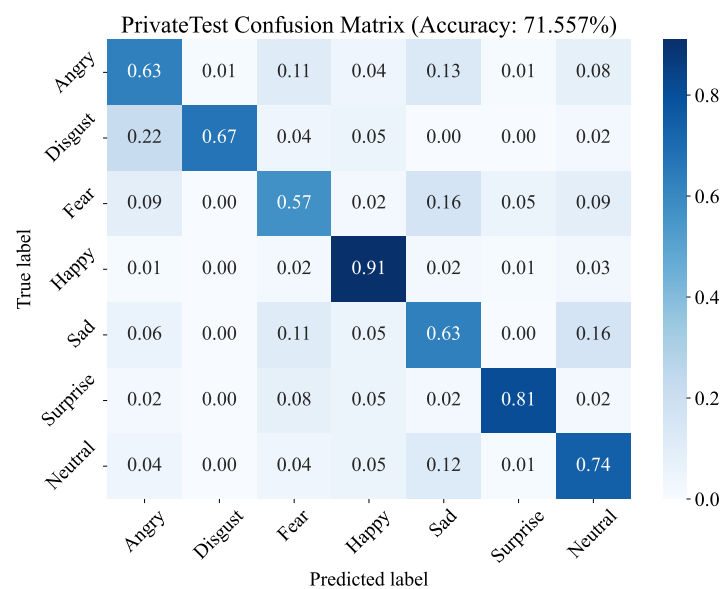
Table 7. Performance evaluation of well-known classification networks and our method on FER2013.

Methods	ACC (%)	Parameters (10^6)
AlexNet [13]	61.1 ± 0.9	63.1
HoG + CNN [25]	63.8 ± 1.3	22.3
VGG [11]	67.4 ± 1.6	100.6
FaceLiveNet [26]	68.0 ± 1.1	69.3
SHNN [27]	69.7 ± 2.3	77.1
EmoNeXt [28]	72.1 ± 0.9	93.6
ResMaskingNet [29]	74.3 ± 1.4	140.3
Our Method	71.1 ± 0.7	28.1

Table 8. Performance evaluation of well-known classification networks and our method on RAF.

Methods	ACC (%)	Parameters (10^6)
DAFL [30]	87.7 ± 0.8	46.9
IF-GAN [15]	88.3 ± 0.9	93.1
PSR [31]	89.0 ± 0.7	55.7
DAN [32]	89.7 ± 1.1	63.6
MRAN [33]	90.0 ± 0.4	77.1
DDAMFN++ [34]	91.4 ± 0.6	84.3
S2D [35]	92.0 ± 0.7	111.3
Our Method	91.0 ± 0.3	28.1

To further substantiate the model's efficacy, this paper employs a confusion matrix to illustrate the recognition accuracy of the half-face-based self-attention weighted model across various expressions. Given the differential impact of half faces on expression classification, employing distinct weights to discern them enhances the performance of expression recognition, as evidenced by Figures 6 and 7. The results demonstrate precise classification of expressions, particularly exhibiting significant recognition capabilities for positive emotions like happiness, aligning with the statistical analysis findings presented earlier.

**Figure 6.** Confusion matrix for FER2013.

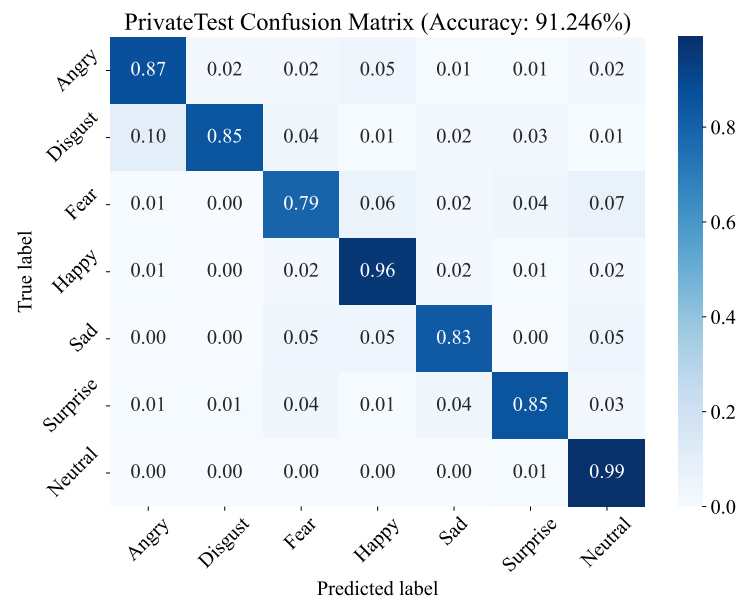


Figure 7. Confusion matrix for RAF.

6. Conclusions

This paper demonstrates the existence of asymmetrical facial expression features through statistical analysis and computer vision techniques. A comprehensive study confirms that the left half-face usually presents more distinguishable expression features.

To address this asymmetry, this paper introduces an innovative half-face self-attention weighted approach for facial expression recognition that handles the different effects of each half-face on expression classification individually. The ability of this technique to assign different weights to the left and right half of the face marks a significant advancement that accommodates the nuances of expression asymmetry. A large number of experiments were conducted on the FER2013 and RAF datasets to validate the effectiveness of the proposed method. The accuracy of this method is close to that of the state-of-the-art recognition models, while using significantly fewer parameters. These findings deepen the current understanding of facial asymmetry in expression recognition and provide a robust framework for future developments in the field.

Author Contributions: S.H. proposed the main idea of the paper and designed the study. X.Y. carried out the numerical simulation and analyzed the examples in the paper. S.H. and X.Y. wrote and revised the paper together. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grants 62276118 and 61772244.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

Acknowledgments: The authors gratefully acknowledge the use of the Third Generation Facial Spontaneous Micro-Expression Database provided by Su-Jing Wang, Institute of Psychology, 16, Lincui Road, Chaoyang District 100101, Beijing, China.

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

References

1. Martini, M.; Bufalari, I.; Stazi, M.A.; Aglioti, S.M. Is that me or my twin? Lack of self-face recognition advantage in identical twins. *PLoS ONE* **2015**, *10*, e0120900. [[CrossRef](#)] [[PubMed](#)]
2. Dopson, W.G.; Beckwith, B.E.; Tucker, D.M.; Bullard-Bates, P.C. Asymmetry of facial expression in spontaneous emotion. *Cortex* **1984**, *20*, 243–251. [[CrossRef](#)] [[PubMed](#)]
3. Liu, Y.; Dai, W.; Fang, F.; Chen, Y.; Huang, R.; Wang, R.; Wan, B. Dynamic multi-channel metric network for joint pose-aware and identity-invariant facial expression recognition. *Inf. Sci.* **2021**, *578*, 195–213. [[CrossRef](#)]
4. Wang, K.; Peng, X.; Yang, J.; Meng, D.; Qiao, Y. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Trans. Image Process.* **2020**, *29*, 4057–4069. [[CrossRef](#)] [[PubMed](#)]
5. Cootes, T.F.; Edwards, G.J.; Taylor, C.J. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *23*, 681–685. [[CrossRef](#)]
6. Ojala, T.; Pietikäinen, M.; Harwood, D. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognit.* **1996**, *29*, 51–59. [[CrossRef](#)]
7. Mountrakis, G.; Im, J.; Ogole, C. Support vector machines in remote sensing: A review. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 247–259. [[CrossRef](#)]
8. Zhang, K.; Huang, Y.; Du, Y.; Wang, L. Facial expression recognition based on deep evolutionary spatial-temporal networks. *IEEE Trans. Image Process.* **2017**, *26*, 4193–4203. [[CrossRef](#)] [[PubMed](#)]
9. Mehta, D.; Siddiqui, M.F.H.; Javaid, A.Y. Recognition of emotion intensities using machine learning algorithms: A comparative study. *Sensors* **2019**, *19*, 1897. [[CrossRef](#)] [[PubMed](#)]
10. 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.
11. Cheng, S.; Zhou, G. Facial expression recognition method based on improved VGG convolutional neural network. *Int. J. Pattern Recognit. Artif. Intell.* **2020**, *34*, 2056003. [[CrossRef](#)]
12. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
13. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*. [[CrossRef](#)]
14. Sun, W.; Zhao, H.; Jin, Z. A visual attention based ROI detection method for facial expression recognition. *Neurocomputing* **2018**, *296*, 12–22. [[CrossRef](#)]
15. Cai, J.; Meng, Z.; Khan, A.S.; O'Reilly, J.; Li, Z.; Han, S.; Tong, Y. Identity-free facial expression recognition using conditional generative adversarial network. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 1344–1348.
16. Cadieu, C.F.; Hong, H.; Yamins, D.L.; Pinto, N.; Ardila, D.; Solomon, E.A.; Majaj, N.J.; DiCarlo, J.J. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.* **2014**, *10*, e1003963. [[CrossRef](#)]
17. Baek, S.; Song, M.; Jang, J.; Kim, G.; Paik, S.B. Spontaneous generation of face recognition in untrained deep neural networks. *Biorxiv* **2019**, 857466.
18. Wen, H.; Shi, J.; Zhang, Y.; Lu, K.H.; Cao, J.; Liu, Z. Neural encoding and decoding with deep learning for dynamic natural vision. *Cereb. Cortex* **2018**, *28*, 4136–4160. [[CrossRef](#)] [[PubMed](#)]
19. 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.
20. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
21. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
22. Wang, K.; Peng, X.; Yang, J.; Lu, S.; Qiao, Y. Suppressing uncertainties for large-scale facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6897–6906.
23. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.H.; et al. Challenges in representation learning: A report on three machine learning contests. In Proceedings of the Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Republic of Korea, 3–7 November 2013; Proceedings, Part III 20; Springer: Berlin/Heidelberg, Germany, 2013; pp. 117–124.
24. Zhang, F.; Zhang, T.; Mao, Q.; Xu, C. Geometry guided pose-invariant facial expression recognition. *IEEE Trans. Image Process.* **2020**, *29*, 4445–4460. [[CrossRef](#)] [[PubMed](#)]
25. Zeng, G.; Zhou, J.; Jia, X.; Xie, W.; Shen, L. Hand-crafted feature guided deep learning for facial expression recognition. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 423–430.

26. 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 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 3507–3512.
27. Miao, S.; Xu, H.; Han, Z.; Zhu, Y. Recognizing facial expressions using a shallow convolutional neural network. *IEEE Access* **2019**, *7*, 78000–78011. [[CrossRef](#)]
28. El Boudouri, Y.; Bohi, A. EmoNeXt: An Adapted ConvNeXt for Facial Emotion Recognition. In Proceedings of the 2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSp), Poitiers, France, 27–29 September 2023; pp. 1–6.
29. Pham, L.; Vu, T.H.; Tran, T.A. Facial expression recognition using residual masking network. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 4513–4519.
30. Farzaneh, A.H.; Qi, X. Facial expression recognition in the wild via deep attentive center loss. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 2402–2411.
31. Vo, T.H.; Lee, G.S.; Yang, H.J.; Kim, S.H. Pyramid with super resolution for in-the-wild facial expression recognition. *IEEE Access* **2020**, *8*, 131988–132001. [[CrossRef](#)]
32. Wen, Z.; Lin, W.; Wang, T.; Xu, G. Distract your attention: Multi-head cross attention network for facial expression recognition. *Biomimetics* **2023**, *8*, 199. [[CrossRef](#)] [[PubMed](#)]
33. Chen, D.; Wen, G.; Li, H.; Chen, R.; Li, C. Multi-relations aware network for in-the-wild facial expression recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 3848–3859. [[CrossRef](#)]
34. Zhang, S.; Zhang, Y.; Zhang, Y.; Wang, Y.; Song, Z. A Dual-Direction Attention Mixed Feature Network for Facial Expression Recognition. *Electronics* **2023**, *12*, 3595. [[CrossRef](#)]
35. Chen, Y.; Li, J.; Shan, S.; Wang, M.; Hong, R. From Static to Dynamic: Adapting Landmark-Aware Image Models for Facial Expression Recognition in Videos. *arXiv* **2023**, arXiv:2312.05447.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.