electronics

MDPI

Article

Improved Facial Expression Recognition Algorithm Based on Local Feature Enhancement and Global Information Association

Zixuan Chen [1], Lingyu Yan [1,*], Hairu Wang [1] and Bogdan Adamyk [2]

[1] School of Computer Science, Hubei University of Technology, Wuhan 430068, China; chenzixuan@hbut.edu.cn (Z.C.); wanghr1086@hbut.edu.cn (H.W.)

[2] Aston Business School, Aston University, Birmingham B4 7ET, UK; b.adamyk@aston.ac.uk

* Correspondence: yanlingyu@hbut.edu.cn

Abstract: Facial expression recognition is the key area of research in computer vision, enabling intelligent devices to understand human emotions and intentions. However, recognition of facial expressions in natural scenes presents challenges due to environmental factors like occlusion and pose variations. To address this, we propose a novel approach that combines local feature enhancement and global information correlation. This method allows the model to learn both local and global facial features along with contextual information. By enhancing salient local features and exploring multi-scale facial expression features, our model effectively mitigates the impact of occlusion and pose variations, improving recognition accuracy. Experimental results demonstrate that our adapted model outperforms alternative algorithms in recognizing facial expressions under challenging environments, achieving recognition accuracies of 85.07% and 99.35% on the RAF-DB and CK+ datasets, respectively.

Keywords: face expression recognition; deep learning; attention mechanism; global information association

check for updates

Citation: Chen, Z.; Yan, L.; Wang, H.; Adamyk, B. Improved Facial Expression Recognition Algorithm Based on Local Feature Enhancement and Global Information Association. *Electronics* **2024**, *13*, 2813. https://doi.org/10.3390/electronics13142813

Academic Editor: Silvia Liberata Ullo

Received: 14 June 2024
Revised: 14 July 2024
Accepted: 15 July 2024
Published: 17 July 2024

1. Introduction

FER (face expression recognition) is a research field that extends from face recognition [1], with the objective of swiftly and precisely identifying face images and categorizing expressions based on facial movements, ultimately enabling the analysis and inference of the emotional state of the subject. To clarify, our computer technology scientifically analyzes facial images and categorizes them into the seven fundamental expressions: happy, surprised, fear, sad, angry, disgusted, and neutral. This classification is accomplished through a systematic sampling procedure. As the information age progresses, the technology of facial expression recognition (FER) has become an integral part of human society, finding diverse applications in various scenarios. Particularly in the field of human–computer interaction [2,3], intelligent robots employ real-time emotional detection with FER to analyze the emotions of individuals. Based on the detection results, they can then execute tailored behavioral responses, enabling intelligent and interactive communication between humans and computers. In the field of human–computer interaction [4], intelligent robots employ FER for real-time emotion detection of target individuals and implement targeted behaviors accordingly, enhancing intelligent human–machine interaction. In the education sector [5], FER enables teachers to infer students' real-time classroom states and knowledge mastery. In the medical field [6], FER aids physicians in accurately assessing patients' conditions and delivering targeted clinical treatment. In the transportation sector, FER analyzes drivers' expressions, promptly generating warning signals and reminders in case of abnormal situations, thereby helping to prevent traffic accidents [7].

Performing (FER) in natural environments poses numerous challenges to researchers, with occlusion and pose variations being particularly tricky. Occlusion leads to the loss of crucial expression information, while pose variations hinder the effective extraction of facial features, both of which collectively reduce the accuracy and robustness of the FER systems. Additionally, the heavy reliance of existing FER models on facial expression image labels, coupled with issues such as high annotation costs and label inconsistencies in public expression datasets, further limits the improvement of model performance.

To address these challenges, this study proposes an innovative FER algorithm, LSDC-FER, which combines local feature enhancement with global information association strategies. By designing a local feature enhancement module, LSDC-FER can strengthen the extraction and enhancement of locally salient features in facial images, enabling it to capture valuable information even in occluded scenarios. Simultaneously, the introduction of a multi-scale global association module allows the model to comprehensively capture the global information and contextual associations of facial expressions across multiple scales, improving its adaptability to various pose variations.

In terms of implementation, LSDC-FER is built upon the ResNet34 backbone network. It first ensures the quality of input images through data preprocessing. Subsequently, the local feature enhancement module divides the feature map into multiple local regions, applies asymmetric convolution blocks to enhance local features, and fuses them back into the global feature map through residual connections. For the purpose of global information association, the integration of multi-scale convolution strategies with the fused convolutional self-attention mechanism (ACMix) proves to be effective in extracting both global multi-scale information and contextual information. Finally, the enhanced local and global features are fused and fed into subsequent convolutional blocks and fully connected networks for expression classification.

Experimental results demonstrate that the LSDC-FER algorithm performs exceptionally well on datasets such as RAF-DB and CK+, achieving recognition accuracies of 85.07% and 99.35%, respectively, significantly outperforming other comparative algorithms. Notably, LSDC-FER exhibits remarkable robustness in handling occlusion and pose variations.

In subsequent chapters, this paper will delve deeper into the current status and limitations of related work, provide a detailed analysis of the technical details of the LSDC-FER algorithm, comprehensively present experimental results and analysis, summarize research findings, and discuss future research directions. Through these efforts, we aspire to contribute to the advancement of facial expression recognition technology.

2. Related Works

2.1. Expression Recognition Method

In the mid-19th century, international scholars began studying facial expressions, with psychologist Paul Ekman [8] leading the way. Ekman conducted extensive experiments, which ultimately led to the identification and classification of six fundamental human expressions: happiness, surprise, fear, sadness, anger, and disgust. These expressions can be seen in Figure 1. Ekman also introduced the Facial Action Coding System (FACS) [9]. The Facial Action Coding System (FACS) divides the face into 46 distinct facial movement units and describes facial expressions by utilizing the combined information derived from these facial action units. Many scholars studying facial expressions accept and base their research on this system. The progression of face expression recognition can generally be categorized into two main types: feature-based face expression recognition methods and deep learning-based face expression recognition methods.

Currently, there are three common types of facial expression recognition methods. The first is the feature-designed expression recognition method, a traditional classification method that predesigns manual features and extracts effective expression information. The classifier has a significant impact on the accuracy of expression recognition, and commonly used classifiers include SVM [10], AdaBoost [11], and K-Means [12]. The second is the geometry-based expression recognition method, which is based on FACS. In 1995, Cootes et al. [13] proposed an active shape model based on statistical learning, which detects the outline of the face and extracts facial feature information to allow the model to more comprehensively extract facial expression features. Matthews et al. [14] proposed an improved algorithm, the active appearance model, based on the active shape model, which enhances the model's ability to detect face contours and locate facial features. Setyati et al. [15] proposed an active shape model combined with a radial deviation function network, which realizes facial expression classification through face reconstruction. Han et al. [16] proposed the face mesh transformation method, which extracts local action units related to expressions through mesh edge feature extraction, achieving an overall accuracy rate of 94.96% in the CK dataset. Finally, there is the texture-based expression recognition method, which presents facial expression feature information by statistically calculating the pixel grayscale distribution of local areas of the facial image. Most of the research methods adopted by researchers are based on classic algorithms such as Local Binary Pattern (LBP) and Gabor wavelet transform. Fu Xiaofeng [17] proposed a multi-scale center binary method based on LBP, which uses the comparison of neighboring point pairs and center pixel weighting to reduce histogram dimensions and introduces an improved LBP sign function and multi-scale to solve the noise sensitivity problem of the LBP operator in facial expression recognition, enabling the model to achieve good classification results. Bashyal et al. [18] extracted facial expression features using 18 Gabor filters and used principal component analysis for data dimension reduction. Finally, they combined vector quantization learning to significantly improve the recognition performance of the algorithm on the Jaffe dataset. Zhang et al. [19] proposed a facial expression recognition algorithm based on Gabor wavelet transform, LBP, and Local Phase Quantization (LPQ). They used Gabor filters to extract facial image features from multiple angles and scales to capture significant facial expression features. The extracted images were encoded using LBP and LPQ operators, and principal component analysis was used to reduce the dimensions of the fused features of the transformed LBP and LPQ operators. Zhang Liang et al. [20] proposed a recognition algorithm based on Gabor wavelet transform and fused gradient feature LBP, which combines the facial region features extracted by the improved LBP operator with Gabor wavelet transform features through weighted fusion, making facial expression features more prominent and achieving good algorithm classification results.

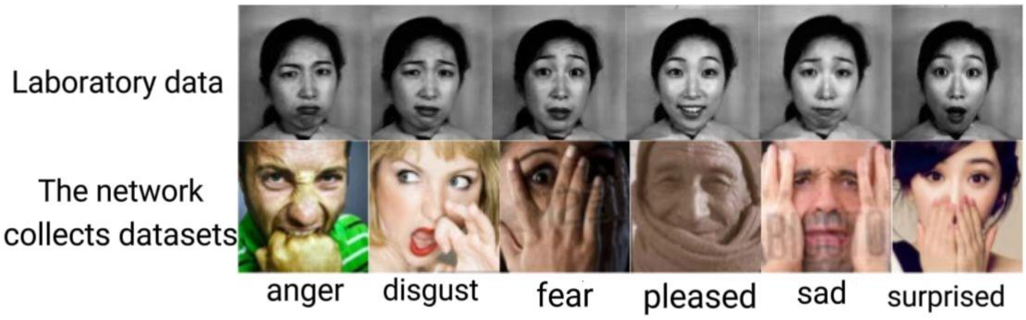


Figure 1. Six types of samples of facial expression data.

Traditional algorithms do not rely on device computing power and are easy to implement, but they do rely on predesigned manual features, and feature extraction and expression classification cannot be optimized together. Additionally, traditional algorithms have weaker robustness and generalization ability in complex scenarios and are difficult to train on large-scale datasets, so their practicality is relatively poor compared to deep learning algorithms.

Yang et al. [21] proposed de-expression residual learning, DeRL, which regards an expression as a combination of expression elements and non-expression elements. The model adversarial network is used to input the model into a unified generation of neutral expression images, and the middle layer information of expression elements is saved by learning adversarial generation network to achieve the expression classification. In a similar vein to DeRL, Ruan et al. [22] introduced a novel feature structure and reconstruction approach called FDRL for face expression recognition. Unlike DeRL, FDRL initially decomposes the expression features generated by the backbone network to derive a collection of potential perceptual features related to facial action units. Subsequently, the model learns the weight of each feature and the weight of the relationship between groups of features to facilitate the reconstruction of face expression features. The weights of the features and their relationships are determined with respect to their importance. The model is able to capture valid expression features. The experiments show that FDRL can achieve a good face expression recognition both in a controlled environment and the natural one. Zhao et al. [23] designed MA-Net, an attention network based on local and global features, for facial occlusion and head pose variation. This network uses Resnet-18 as its backbone, mitigating the interference from occlusion factors. Test results indicate that MA-Net can effectively achieve facial expression recognition in natural scenarios. Wang et al. [24] also acknowledged that occlusion and pose variation hinder facial expression recognition technology and proposed the Region Attention Network (RAN) to adaptively extract effective facial expression features. Additionally, Wang et al. [25] viewed deep learning-based facial expression recognition research from another perspective, arguing that uncertainties arising from low-quality images and non-objective expression image labels seriously mislead the learning of neural networks. They proposed the Self-Cure Network (SCN) to suppress the uncertainties faced by facial expression recognition. Unlike suppressing uncertainties, Zhang et al. [26] also proposed a new solution for the uncertainty problem. They encouraged the model to learn more precise uncertainty values through the loss function, helping the model learn labels for ambiguous expression images from mixed images.

Numerous studies indicate that, in comparison to traditional expression recognition methods, deep learning-based approaches exhibit superior recognition capabilities. They are adept at learning intricate expression patterns and contextual information, and typically demonstrate high accuracy, particularly when trained on large-scale datasets and equipped with a sufficiently deep network structure. Moreover, these methods also demonstrate a certain level of robustness to image variations and noise. Even in the presence of image noise, deformations, and other distortions, deep learning models can effectively recognize facial expressions. They possess stronger generalization capabilities and capture a wider range of expression variations and sample diversity through extensive training data. This enables the method to achieve favorable recognition outcomes on unseen data.

The uniqueness and significance of this study lie in the application of the global information association module and local feature enhancement. Firstly, by introducing a multi-scale global association module, this study achieves deep excavation and fusion of global facial expression information. Furthermore, the integration of the fused convolutional self-attention mechanism (ACMix) dynamically captures the contextual associations within facial expressions. The effective utilization of the ACMix mechanism enables the model to intelligently allocate attention, focusing on key facial regions during expression recognition, thereby significantly enhancing the robustness and accuracy of the model in handling complex natural scenarios.

Echoing the global information association, this study also conducts meticulous processing in local feature enhancement. Through fine-grained segmentation of feature maps and the introduction of asymmetric convolution blocks, the local feature enhancement module precisely captures and effectively enhances the crucial local features of facial expressions. This meticulous treatment not only improves the expressiveness of local features but also seamlessly fuses them with the global feature map through residual connections, achieving complementarity and enhancement between local and global information. Table 1 shows a comparison of the advantages and disadvantages of the traditional method and ours.

The proposed method in this paper stands out for its ability to integrate global and local facial features, dynamically model facial contexts across multiple scales, and enhance salient local features. This unique combination enables robust facial expression recognition, even under challenging conditions like occlusion and pose variations.

Table 1. Comparison of the advantages and disadvantages of the traditional method and the method in this paper.

Model/Methodology	Advantages	Disadvantages
LSDC-FER	Global and local information fusion: the global and local features are effectively fused, which improves the accuracy and robustness of recognition. Multi-scale feature extraction: the generalization ability of the model is enhanced. Dynamic context modeling: the ACMix mechanism can dynamically capture the contextual association of facial expressions, which improves the ability of the model to handle complex scenes. Robust: excellent in handling complex situations such as occlusion and pose changes.	Computational complexity and resource requirements are high. Several parameters in the model need to be carefully tuned for optimal performance.
Existing methods (e.g., De-Ken, Pugh-Ken, etc.)	High computational efficiency for real-time applications. The model is simple and easy to implement and deploy. The technology is mature.	Relying too much on local features degrades performance in complex scenarios. Limited generalization ability. Lack of dynamic modeling. Sensitive to parameters.

2.2. *Face Expression Dataset*

A pivotal aspect of facial expression recognition based on deep learning is the selection of the dataset. Ideally, the dataset ought to comprise facial expression images representing a wide array of races, cultures, and environments. In this section, we use the public emoji dataset in the proposed algorithm. According to the dataset collection source, the dataset can be divided into laboratory dataset and network collection dataset, as shown in Table 2.

Table 2. Sample distribution of RAF-DB dataset.

Category	Table	Training Set	Test Set
Neutral	0	2524	680
Pleased	1	4772	1185
Sad	2	1982	478
Surprised	3	1290	329
Fear	4	281	74
Disgust	5	717	478

The RAF-DB dataset [27] is a publicly available dataset designed for research in facial expression recognition and sentiment analysis. It includes face images sourced from real-world scenes on the Internet and covers diverse factors such as age, gender, and skin color. The dataset contains 29,672 images in total. Furthermore, the dataset was annotated by 40 professional expression annotators, incorporating both basic expression and composite expression datasets. The training set consisted of 12,271 samples, while the test set contained 3068 samples. The detailed distribution is given in Table 3.

Table 3. Sample distribution of FER-2013 dataset.

Category	Table	Training Set	Test Set
Neutral	0	4965	626
Pleased	1	7215	879
Sad	2	4830	594
Surprised	3	3171	416
Fear	4	4097	528
Disgusted	5	436	55

FER-2013 is a publicly available emoji dataset, that consists of a substantial collection of 35,887 images obtained from the Internet, as depicted in Table 3. Unlike RAF-DB, the FER-2013 dataset maintains a fixed image size of 48 × 48 pixels. However, there is a challenge where the image-to-label correspondence is inconsistent. Moreover, the presence of non-face images within the FER-2013 dataset poses difficulties in achieving high accuracy for the algorithm.

The CK+ dataset [28] is among the datasets commonly used for evaluating expression recognition algorithms. It is an enhanced version of the original Cohn Kanade dataset, refined by subsequent researchers. The CK+ dataset contains both images and videos, although only images are utilized for training and testing purposes in this experiment. It is noteworthy to mention that the CK+ dataset is a collection of facial expressions gathered in a controlled laboratory environment.

3. Methods

3.1. Technical Principles

Facial expression recognition often faces problems such as occlusion and posture change in natural scene applications [29]. The algorithm struggles to capture expression information from key parts; therefore, it has some impact on the accuracy of facial expression recognition [30]. Previous studies were mainly focused on detecting and addressing affected local areas, which are complex, deeper in structure, and resource-intensive [31]. Therefore, our proposal involves developing an adaptive network model called GCLENet, which aims to enhance local features and integrate global information. This approach effectively mitigates the impact of pose transformations and occlusions on facial expression recognition, leading to improved accuracy and robustness in real-world scenarios. By combining both local and global features, GCLENet addresses the challenges associated with variations in facial expressions, ultimately enhancing the overall performance of the recognition system.

As illustrated in Figure 2, the backbone network of GCLENet is based on ResNet-34 [32], which, in its structural design, can be specifically divided into two primary modules: the local feature enhancement module and the multi-scale global association module. After the expression image is initially processed using the convolution operation of the first two convolution blocks of ResNet-34, it undergoes further processing by both the local feature enhancement module (LFA) and the multi-scale global association module (MGC) [33,34]. The purpose of this processing is to extract local features, global features, and global context information. Specifically, within the local feature enhancement branch module, the middle-level face feature map is initially segmented into multiple non-overlapping local feature maps, following the spatial horizontal and vertical directions. Convolution blocks are subsequently utilized to capture the unique local features of the feature-level face. In the multi-scale global association module, the feature map initially learns the global multi-scale information of the face by passing through the multi-scale module. Next, the fusion convolution (FCM) derived from the attention mechanism ACMix [26] is introduced. This approach aims at effectively extracting the global context information of the face while simultaneously convolving the global multi-scale feature map [26,35]. The local features obtained at the feature level are then fused with the global features, which contain fusion context information. Finally, these fusion features are fed into the subsequent convolution group and fully connected network for accomplishing face expression classification.

In this section, this paper will present the facial expression recognition method that enhances the integration of global information with local features. The method's efficacy will be demonstrated with practical experiments, to confirm its exceptional performance in accurately identifying and categorizing facial expressions.

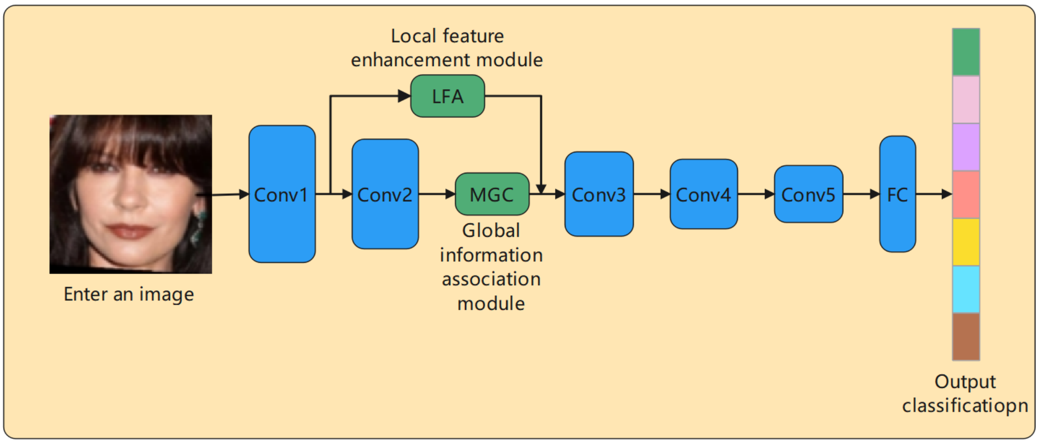

Local feature enhancement module
LFA
Conv1
Conv2
MGC
Global information association module
Conv3
Conv4
Conv5
FC
Enter an image
Output classificatiopn

Figure 2. The structure diagram of GCLENet.

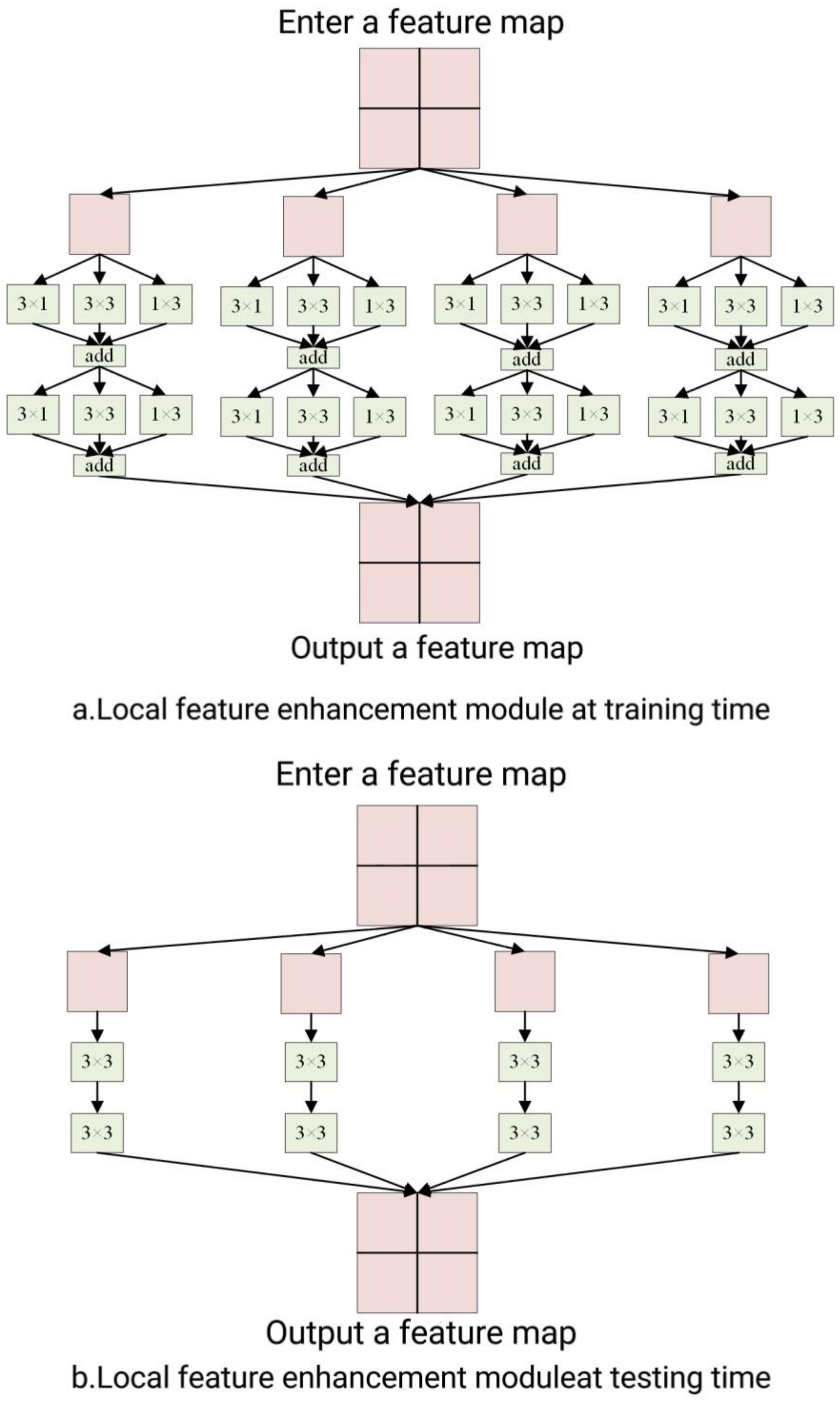
3.2. Local Feature Enhancement Module

Local features play a crucial role in recognizing facial expressions, as they offer more precise characteristics that describe facial expression information, as noted in [36]. Previous approaches mainly extracted local features by cropping specific facial regions. However, challenges such as occlusions and pose variations often result in coverage and deformation of these key regions, making it challenging to implement such algorithms in natural environments. For this purpose, GCLENet introduces the local feature enhancement module. As illustrated in Figure 3, LFA simplifies the process by eliminating the need for complex face detection and key area clipping. Instead, it focuses on area segmentation and local operations directly on the global feature map. This approach ensures the extracted local features of the face are at the feature level. The multi-layer convolution further enhances the expressive capability of these local facial features, effectively mitigating the interference from occlusion and pose variations on the global scale. Additionally, subsequent convolution operations with residuals are incorporated after local feature extraction, enabling the model to learn more comprehensive facial representations.

A face image with the dimensions of 3 × 224 × 224 is sent to the input of the backbone network. The feature map F is generated after 7 × 7 convolution and maximum pooling of the first layer of the base block, which contains rich shallow detail features. The parameters are as follows: $F \in R^{C \times H \times W}$, $C = 64$, $H = W = 56$. The results reveal that the crux of recognizing facial expressions lies in the eyes and mouth, and it is imperative to also consider the inherent attribute factors of faces. To begin, the feature map F is divided into four feature submaps along the spatial direction F_i, thereinto $i \in \{1, 2, 3, 4\}$. The facial image is divided into four regions of interest, each represented as a feature subplot with the size of 28 × 28. These subplots focus on crucial facial areas. To enhance the local features of the face, we employ two sets of asymmetric convolutions. The asymmetric convolution block is composed of three convolution kernels: 1 × 3, 3 × 3, and 3 × 1. This configuration is mathematically represented as

$$F_i' = conv_{1\times3}(F_i) + conv_{3\times3}(F_i) + conv_{3\times1}(F_i) \tag{1}$$

This enhances the ability to represent local features of facial expressions. Following two asymmetric convolutions, the four feature submaps are reconnected in their original sequence. Then, the local feature maps are merged into the global features as residuals. The LFA module structure during the training process is shown in Figure 3a.



a.Local feature enhancement module at training time

b.Local feature enhancement moduleat testing time

Figure 3. The structural design of the module for local features enhancement.

It is important to note that the model introduced in this paper utilizes asymmetric convolutional blocks exclusively for local enhancement during the training process. However, during the testing phase, these asymmetric convolution blocks are substituted with standard 3 × 3 convolutions. The weight parameters of 1 × 3 and 3 × 1 are loaded into the central position of the 3 × 3 convolution kernel, and the LFA module structure during the test phase is shown in Figure 3b.

Facial expressions are commonly generated by the movement and alteration of local areas, and LFA allows the model to analyze specific parts of these expressions, enhancing its ability to comprehend their inherent nature. Furthermore, facial expressions are manifested differently across various scenarios. The incorporation of LFA effectively filters out occlusion and pose variation information, ultimately enhancing the robustness of the model.

3.3. Multi-Scale Global Association Module

In the classification of facial expressions, affected by the interclass similarity of facial expressions, most algorithms ignore the importance of global features and pay more attention to local features. Previous studies revealed the significance of global features in recognizing facial expressions under natural scenes. Therefore, we have designed a module based on ACMix that is capable of extracting global multi-scale features and fusing global context information to enhance recognition. As illustrated in Figure 4, MGC learns the global feature information of facial expressions through a two-stage structure. In the first stage, we map the intermediate face features generated by multiple convolutions. For this purpose, F is evenly divided into n feature subgraphs along the direction of the channel F_i, thereinto $i \in \{1, 2, \ldots, n\}$. Except that the first feature subgraph F_1 uses 1×1 convolutions, the remaining subgraphs were extracted by convolution kernels with different receptive fields. The output subgraph Y_i can be represented as follows:

$$Y_i = \begin{cases} o(F_i) & i = 1 \\ \underbrace{K(\ldots K(F_i)\ldots)}_{(i-1)K(\bullet)} & 1 < i \leq n \end{cases} \tag{2}$$

where $o(\cdot)$ stands for 1×1 convolution and $K(\cdot)$ stands for 3×3 convolution. Lastly, the generated feature submap is concatenated along the channel direction to obtain a comprehensive global multi-scale feature map $Y = concat(Y_1, Y_2, \ldots, Y_n)$. Furthermore, the value of n has a significant impact on the global features learned by the model. Specifically, as the value of n increases, the global features captured by the model become richer and more comprehensive, but this leads to an increase in operational cost. Therefore, a balance between feature learning and operational cost must be struck. We have set the value to 4, that is, in the same convolutional layer, respectively, the receptive fields of 1×1, 3×3, 5×5, 7×7 convolution check for convolution of four sets of feature subgraphs, as shown in Figure 4.

The introduction of multi-scale features can indeed empower the model to learn stronger and more robust global expression information [37]. But from the FACS theory, facial expressions are jointly described by facial action units. Learning the association between expression features aids the model's recognition of facial expressions, improving its overall performance [38,39]. To accomplish this, we introduce ACMix, a two-branch fusion convolutional self-attention mechanism that allows the model to effectively learn the global context information of facial expressions while simultaneously continuing to learn facial features, as depicted in Figure 5. The output in traditional 3×3 convolutions can be represented as the summation of feature maps, which are computed using pixels as the basis:

$$\begin{cases} g_{i,j}^{(p,q)} = \sum\limits_{p,q} K_{p,q} f_{i+p-\lfloor k/2 \rfloor, j+q-\lfloor k/2 \rfloor} \\ g_{i,j} = \sum\limits_{p.q} g_{i,j}^{(p,q)} \end{cases} \tag{3}$$

where $f_{i,j}$ and $g_{i,j}$ denote the pixel positions of the input and output feature maps, respectively, and $K_{p,q}$ represents the weight of the traditional convolution kernel at position p, q. The top expression in Equation (3) can be simplified by the shift operation:

$$g_{i,j}^{(p,q)} = Shift(K_{p,q} f_{i,j}, p - \lfloor k/2 \rfloor, q - \lfloor k/2 \rfloor) \tag{4}$$

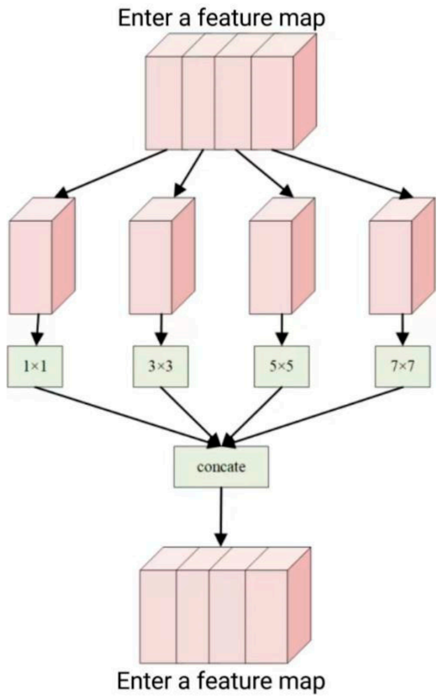


Figure 4. The diagram of multi-scale global feature extraction.

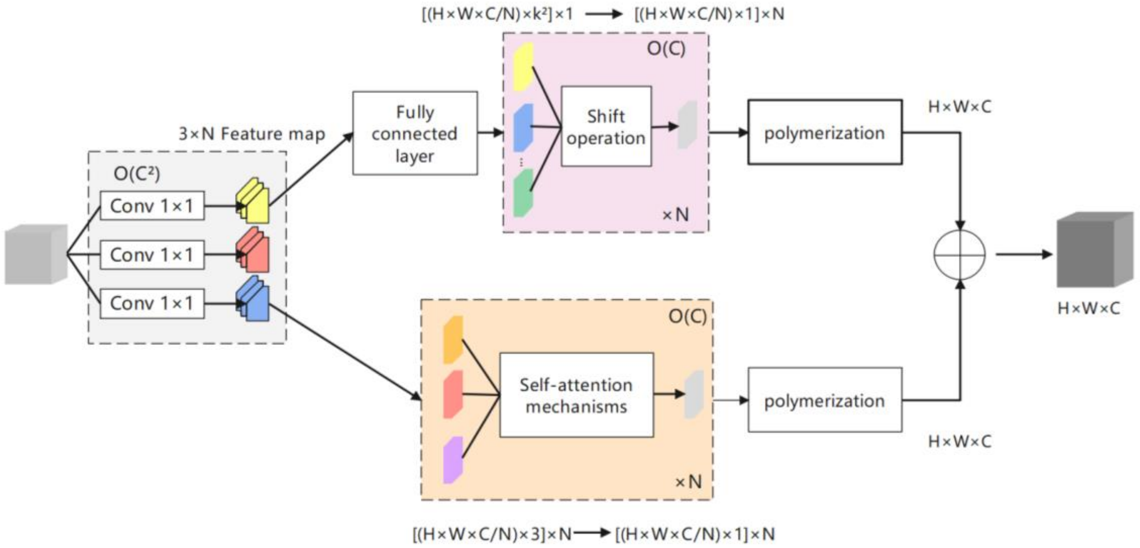


Figure 5. The convolutional attention structure diagram is fused in the module of multi-scale global information association.

The last two parameters represent the horizontal and vertical displacements of $f_{i,j}$. Similarly, the essential concept of the self-attention mechanism can be articulated through the following formula:

$$A(W_q f_{i,j}, W_k f_{a,b}) \cdot W_v f_{a,b} = W_v f_{a,b} \cdot softmax_{N_k(i,j)}(\frac{W_q f_{i,j}, W_k f_{a,b}}{\sqrt{d}}) \tag{5}$$

where W_q, W_k, and W_v denote the parameter matrices for query, key, and value, respectively; $N_k(i,j)$ signifies the local region of the pixel; and d represents the feature dimension of $W_q f_{i,j}$. By harnessing the strengths of both convolutional neural networks and self-attention mechanisms, we can leverage the shared characteristics of feature maps within the same module. In our specific implementation, given a multi-scale facial feature map

F, we initially subject the feature map to pre-projection. This process typically entails applying three 1 × 1 convolutions to transform the individual feature maps into n separate groups, ultimately resulting in a diverse set of intermediate features. The feature map is subsequently reconstructed via the fully connected layer and goes through the processes of biasing and aggregation. This process effectively produces a feature map that serves as an equivalent to the convolutional kernel K for the subsequent convolutions. This combined approach allows us to effectively leverage the benefits of both convolution and self-attention mechanisms, enhancing the representation and processing capabilities of our model. Secondly, by employing 1 × 1 convolutions, the feature projections for query, key, and value can be obtained simultaneously. The channels for query and key are then reshaped into vectors using the self-attention mechanism. A similarity matrix is generated via matrix point multiplication, which is subsequently weighted with the value. Lastly, the outputs from the two branches are adaptively weighted through parameter control to obtain the final global feature map.

Therefore, compared to other algorithms, this network extracts a broader range of features with diverse receptive fields through multi-scale modules. Furthermore, it integrates fusion convolutional self-attention to capture the internal correlation among expression features, effectively mitigating the interference caused by occlusion and posture changes on the network. This ultimately enhances the performance and robustness of the algorithm.

The global information association module leverages multi-scale feature extraction and the self-attention mechanism (ACMix) to achieve a deep understanding and dynamic modeling of global information of facial expressions, effectively enhancing the model's recognition capabilities in complex scenarios [40]. Simultaneously, the local feature enhancement module significantly boosts the extraction and expression capabilities of local features through fine-grained segmentation of feature maps and the introduction of asymmetric convolution blocks [41]. Furthermore, it integrates these local features with the global feature map through residual connections, creating a complementary and enhanced feature representation [42,43].

The design of these two modules is grounded on solid scientific theoretical foundations, not only addressing long-standing challenges in the FER field but also enhancing the efficiency and value of practical applications of the algorithm.

4. Experiments and Analysis

4.1. Implementation Details

Experiments conducted on 64-bit Ubuntu 16 systems utilized GeForce GTX 1070Ti graphics for acceleration. The implementation involved two sections: training and verification. Typically, pre-training with a relevant dataset is utilized for face expression recognition. Previous studies showed that pre-training on a large-scale face dataset can enhance the results with fine-tuning for face expression recognition. However, large-scale pre-training may require substantial resources. To address this, we assembled a training set consisting of 50,000 facial expression images from three face expression datasets: CAER-S [44] (see Figure 6), FED-RO [45] (see Figure 7), and AffectNet [46] (see Figure 8). In the training process, we utilized MTCNN for face image detection and cropping before inputting them into the network. The model's parameters were configured as follows: a mini-batch size of 128, an initial learning rate of 0.05, with the learning rate decreased by a factor of 10 every 20 iterations, for a total of 200 training iterations. The loss function utilized in this study is cross-entropy, and the optimizer algorithm employed is stochastic gradient descent. The momentum parameter is set to 0.9, with the intention of accelerating the convergence speed during the optimization process. Meanwhile, the weight decay parameter is set to 0.0005, aiming at preventing model overfitting by slightly penalizing the weights and thereby enhancing the model's generalization ability. During the model validation phase, we incorporated the trained model parameters into the model and set the learning rate equal to 0.01 and performed 100 additional training iterations. The rest of the parameters remained unchanged. This allows us to assess the model's performance with independent validation data.

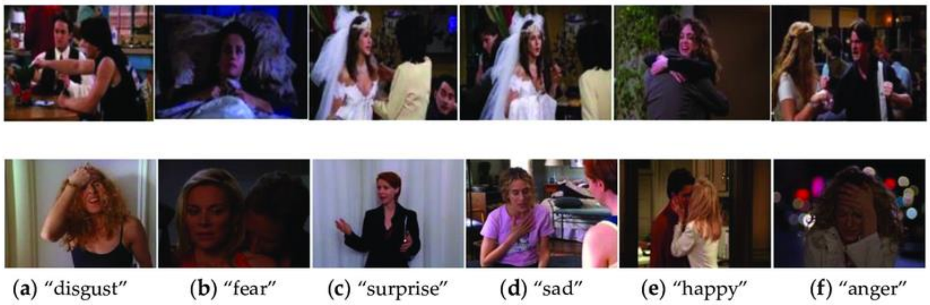


Figure 6. Illustrative frames from the CAER-S dataset.

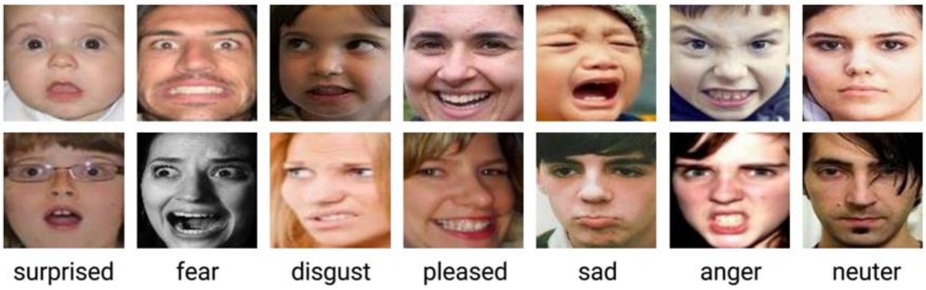


Figure 7. Illustrative frames from the FED-RO dataset.

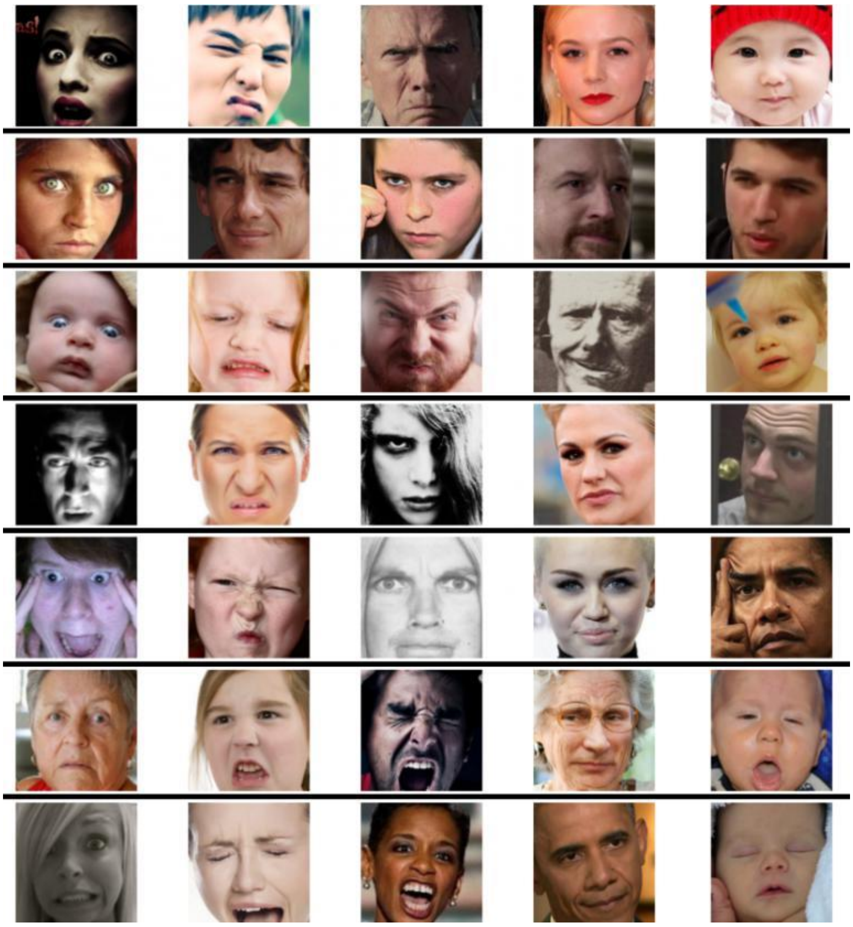


Figure 8. Illustrative frames from the AffectNet dataset.

4.2. Evaluation Indicators

The confusion matrix helps to assess the model's performance by providing insights into its accuracy in classifying different categories. It allows computing the performance metrics, including accuracy, recall, precision, and F1 score, providing a thorough understanding of the model's classification capabilities.

Referring to Table 4, consider a binary classification scenario. The four cells of the confusion matrix represent true positives, false positives, true negatives, and false negatives, respectively. The term "true positive" means the cases where the model correctly identifies a positive example as positive. Conversely, the term "false positive" refers to cases where the model incorrectly classifies a negative example as positive. Similarly, "true negative" refers to cases where the model accurately classifies a negative example as negative, whereas "false negative" indicates cases where the model mistakenly classifies a positive example as negative. The accuracy can be calculated by Formula (6). We used only these metrics to assess the performance of the recognition methods.

$$Accurracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{6}$$

Table 4. Sample confusion matrix diagram.

Confusion Matrix		**True Value**	
		Positive	**Negative**
Predicted value	Positive	TP	FP
	Negative	FN	TN

4.3. Experimental Results

In this section, we present a comparison of the performance of our algorithm with other face expression recognition algorithms, using two distinct datasets, such as RAF-DB and CK+ [47,48]. These datasets are commonly used for evaluating the accuracy and effectiveness of facial expression recognition algorithms. To ensure a fair evaluation of algorithm performance, we adhered to the specific requirements of each algorithm. For the algorithms that require pre-training, we conducted model pre-training using the ImageNet dataset. This process establishes a strong foundation for accurate evaluation. Conversely, for the algorithms that do not require pre-training, we directly assessed their performance without this supplementary preprocessing step. This approach enables a comprehensive and reliable comparison of algorithm performance.

(1) Performance comparison on RAF-DB datasets.

The RAF-DB dataset is composed of two sets: the original face image set and the aligned face image set. Therefore, for performance verification, we can directly use the preprocessed aligned face image set. Figure 9 displays the comparison of accuracy, demonstrating the performance of the algorithms on the dataset under evaluation. The performance of the GCLENet algorithm on RAF-DB is better than the other ones, and its recognition accuracy reaches 85.07%. Compared to DLP-CNN [49], PG-CNN [50], and IRF-CNN [51], GCLENet performed better by 0.94%, 1.8%, and 1.53%, respectively. Hence, the superior performance of GCLENet indicates its ability to effectively address the challenges posed by occlusion and posture changes in face expression recognition in natural scenes [52].

(2) Performance comparison on the CK+ dataset.

On the CK+ emoji dataset, we tested model performance using only 758 images. There is not a precise division between the training set and the test set in the official CK+ dataset. In this paper, we took on the task of partitioning the official CK+ dataset into a training set and a test set, with the ratio of 4:1. This division allows us to conduct a more systematic evaluation of the algorithm performance on this dataset. According to the experimental results depicted in Figure 10, GCLENet attains an accuracy of 99.35% on the CK+ dataset, outperforming the other methods.

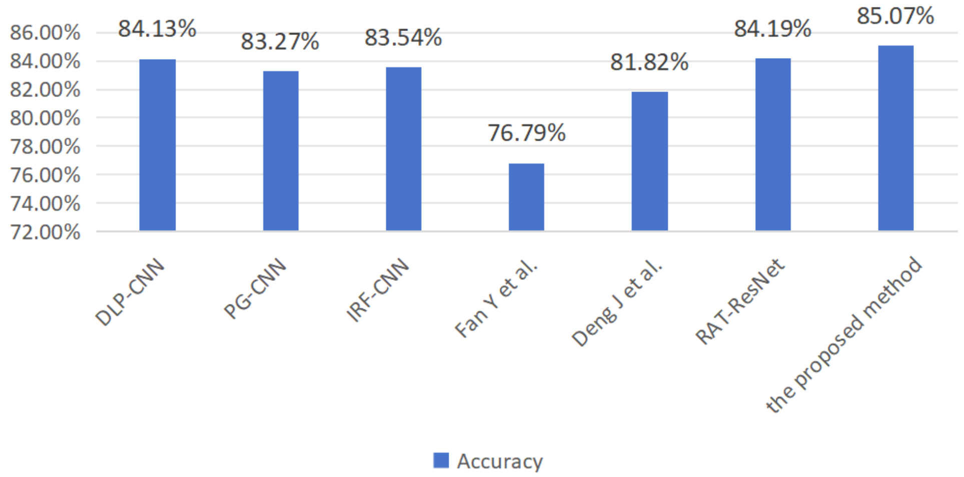


Figure 9. The performance comparison of different algorithms with RAF-DB datasets using the accuracy as a parameter for comparison.

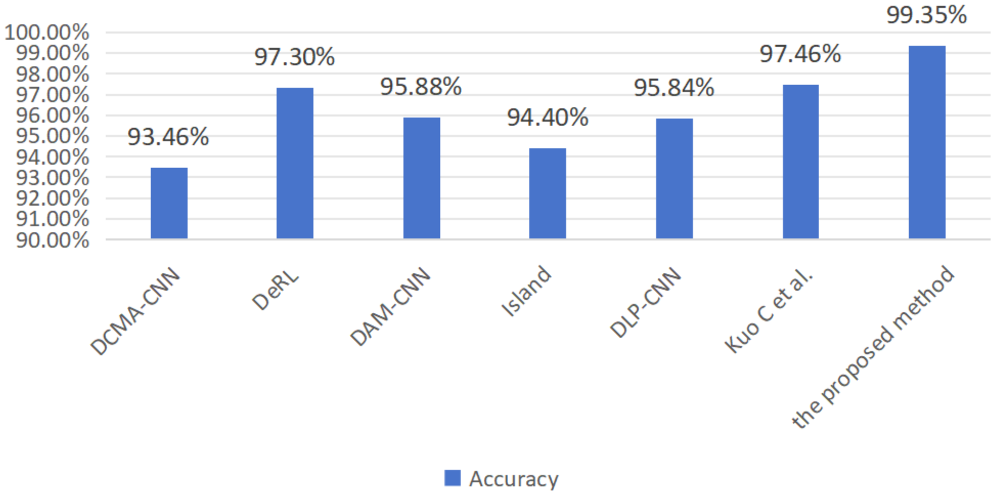


Figure 10. Performance comparison of different algorithms using the CK+ dataset.

4.4. Ablation Experiments

The performance enhancement of the GCLENet network can be ascribed to both the local feature enhancement module and the multi-scale global association module. To gain a comprehensive understanding of the impact of these modules and their internal configurations on face expression recognition in natural scenes, ablation studies were conducted on the RAF-DB and CK+ datasets. These experiments were aimed at evaluating the individual contributions of each module in enhancing the model's recognition performance.

(1) Module ablation experiment

In Table 5, without adopting LFA and MGC, the model achieved accuracy rates of 83.38% and 96.92%, respectively. This is a slightly lower performance. Incorporation of the LFA module alone resulted in significant accuracy improvements of 1.31% on RAF-DB and 1.96% on CK+, respectively. Similarly, integrating the MGC module alone increased the model accuracy by 0.77% on RAF-DB and 1.29% on CK+, respectively. These results emphasize the individual contributions of LFA and MGC in enhancing the model's recognition performance on both datasets. Lastly, the simultaneous utilization of both modules led to accuracy improvements of 1.69% on RAF-DB and 2.43% on CK+ with respect to the basic version.

Table 5. Results of the ablation studies.

LFA	MGC	RAF-DB	CK+
-	-	83.38%	96.92%
√	-	84.69%	98.88%
-	√	84.15%	98.21%
√	√	85.07%	99.35%

(2) Local feature enhancement module internal analysis

The local feature enhancement module plays a crucial role in the GCLENet network, as it enhances local features through feature map segmentation and asymmetric convolution. The selective utilization of these techniques validates the effectiveness of the local feature enhancement mechanism. This targeted extraction of local features significantly enhances the overall performance of the network.

We evaluate four local feature extraction combination strategies here. In the comparative experiments presented in Table 6, four different approaches were employed to extract local features. The first approach is based on traditional 3 × 3 convolutions extracting local features in a conventional manner without segmenting the feature map. The second approach divided the original feature map into four groups and performed 3 × 3 convolution on each segmented feature subgraph to extract local features. The third approach directly applied asymmetric convolution to the feature submap, focusing on extracting local features. Lastly, the local feature enhancement module developed in this paper, was utilized as the fourth approach. The experimental results and performance of these four groups can be found in Table 7. Undoubtedly, the local feature enhancement module exhibits a stronger effect in comparison with the other groups. This can be attributed to the feature map segmentation, which enables the model to more effectively concentrate on the local facial features. Furthermore, the module derives advantages from the influence of the MANet network. Additionally, the four groups of uncovered feature maps align well with the facial expression structure, further enhancing the validity of the local feature enhancement mechanism. At the same time, the use of asymmetric convolution can effectively enhance the extraction of local facial features, which is effective for recognizing facial expressions in natural scenes.

Table 6. Analysis of studies of the local feature enhancement module.

Method	RAF-DB	CK+
Baseline	83.38%	96.92%
Baseline + Asymmetric convolution	83.87%	97.84%
Baseline + Feature map segmentation	84.26%	98.42%
Baseline + Local feature enhancement module	84.69%	98.88%

Table 7. Analysis of studies of the global multi-scale association module.

Method	RAF-DB	CK+
Baseline	83.38%	96.92%
Baseline + Multi-scale modules	83.72%	97.33%
Baseline + Fusion convolutional attention	83.96%	97.58%
Baseline + Global information association module	84.15%	98.21%

(3) An internal analysis of the multi-scale global correlation module

The multi-scale global association module consists of the multi-scale convolution and fusion convolution self-attention. Its purpose is to investigate the effectiveness of incorporating fusion context information into the global multi-scale model. To verify this, four sets of comparative experiments were conducted. In the first set of experiments, the effectiveness of the proposed multi-scale global association module was evaluated. In the second set, the multi-scale convolution was replaced with the 3 × 3 convolution. The third set of experiments involved removing the fusion convolution attention and reusing the 3 × 3 convolution. Finally, the fourth set of experiments eliminated the entire multi-scale global association module while keeping the remaining conditions the same. The results are presented in Table 7.

The results indicate that the performance of the first group of experiments surpasses those of the other groups. It is evident that the multi-scale global association module contributes to the acquisition of diverse facial expression features during model learning. Furthermore, the incorporation of convolutional self-attention highlights the internal correlations among facial expression features, effectively mitigating the influence of occlusion and pose variations on the learning process of the network model.

(4) Analysis of experimental results

To gain further insights into the training process and evaluate the performance of GCLENet, this study utilized the matplotlib library to visualize the pertinent data associated with the training of GCLENet on the RAF-DB dataset. By utilizing data visualization techniques and conducting thorough analysis, a comprehensive understanding of GCLENet's performance was obtained. Figure 11 displays four curves, where the solid dark green line represents the training accuracy curve, and the solid light green line represents the test accuracy curve. On the other hand, the dark green dashed line and the light green dashed line depict the training loss function curve and the test loss function curve, respectively. The horizontal axis represents the model iteration round, while the vertical axis indicates both accuracy and the loss function. It is noteworthy that, for visualization purposes, the scale of the loss function is enlarged by a factor of 30. The figure clearly demonstrates that within the first 30 iterations, the model experienced a rapid increase in both the training recognition accuracy curve and the test recognition accuracy curve, achieving an initial stage of convergence. Over the subsequent 70 iterations, the model reached convergence and sustained a stable accuracy rate of roughly 85%. Eventually, the model attained the final accuracy of 85.07%. Correspondingly, the loss function curve demonstrated an inverse trend compared to the accuracy curve. Notably, the curve for the proposed method displayed a consistent trend without significant fluctuations, indicating that our algorithm successfully converged.

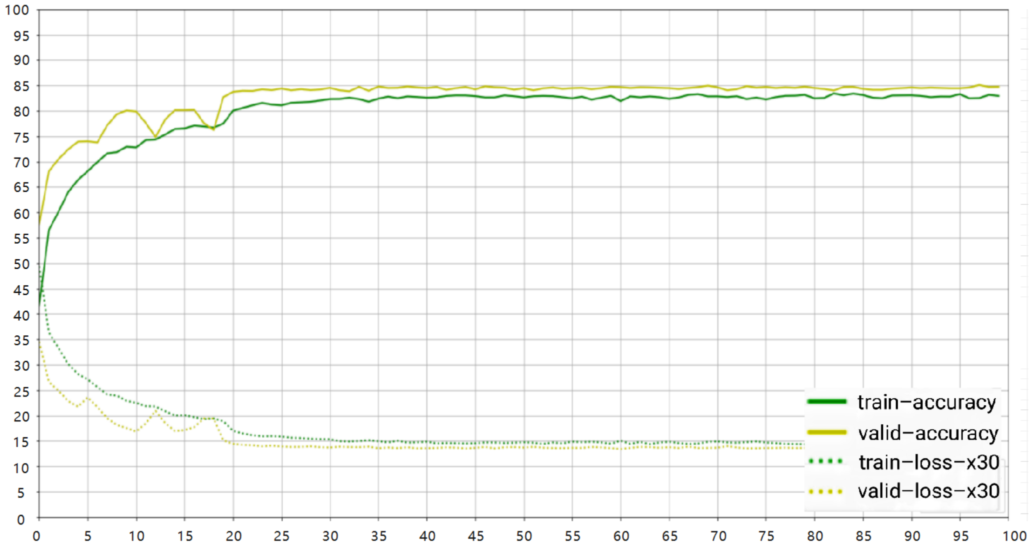
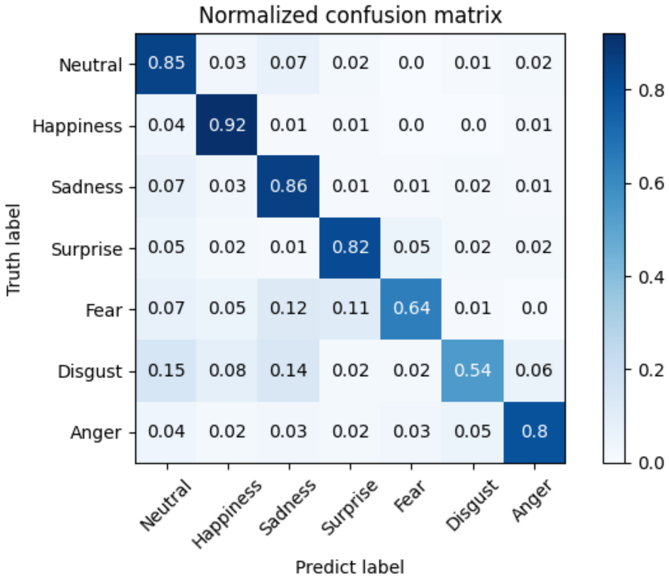


Figure 11. GCLENet iteration diagram of 100 rounds on RAF-DB100.

In addition, to provide a more comprehensive evaluation of the performance of GCLENet on the RAF-DB dataset, we used confusion matrices and gradient cam graphs. These visualizations provide more detailed information and allow for a more accurate description and analysis of the model's recognition performance. By using these additional methods, we aimed to avoid relying solely on accuracy and loss function curves when evaluating GCLENet. According to the description in Figure 12, the confusion matrix shows that the algorithm exhibits a relatively weak ability to recognize two specific facial expressions, such as fear and disgust. The accuracy of identifying these expressions is only 64% and 54%, respectively. These figures are significantly lower than those of other expression categories. This may be attributed to the relatively small number of disgust and anger samples in the RAF-DB expression dataset, as well as the high similarity between expressions caused by occlusion and posture changes. Therefore, distinguishing between these two types of expressions and other classes becomes very challenging, which in turn hinders the learning process of the model.



Truth label \ Predict label	Neutral	Happiness	Sadness	Surprise	Fear	Disgust	Anger
Neutral	0.85	0.03	0.07	0.02	0.0	0.01	0.02
Happiness	0.04	0.92	0.01	0.01	0.0	0.0	0.01
Sadness	0.07	0.03	0.86	0.01	0.01	0.02	0.01
Surprise	0.05	0.02	0.01	0.82	0.05	0.02	0.02
Fear	0.07	0.05	0.12	0.11	0.64	0.01	0.0
Disgust	0.15	0.08	0.14	0.02	0.02	0.54	0.06
Anger	0.04	0.02	0.03	0.02	0.03	0.05	0.8

Figure 12. The confusion matrix of the model in RAF-DB.

As illustrated in Figure 13, through gradient cam visualization, it can be observed that compared to the original ResNet34 network for image processing, GCLENet exhibits a higher accuracy in focusing on the key non-occluded areas within facial images. This indicates that the algorithm in this part can effectively guide the model to focus on facial expressions, thereby reducing the impact of occlusion and pose changes.

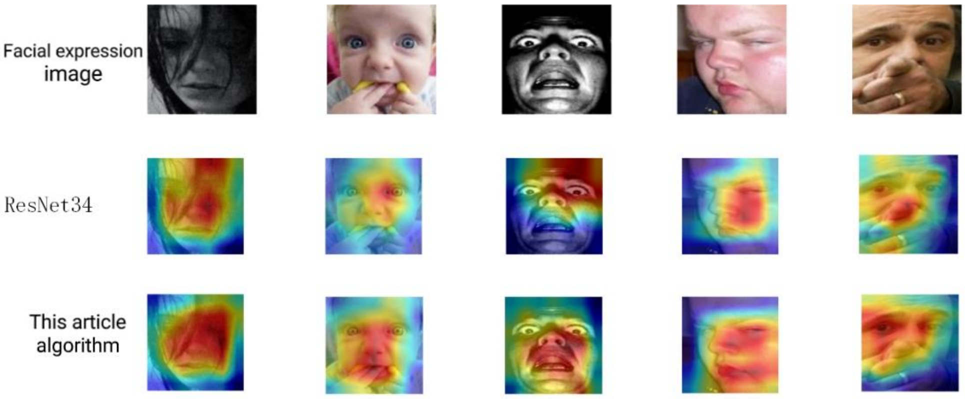


Figure 13. Comparison of the algorithms using some facial expressions.

4.5. Model Complexity Analysis

In this section, we introduce complexity analysis during the testing phase to evaluate the algorithms. As shown in Figure 14, the baseline model ResNet34 achieves 21.29 M of the total parameters calculated by the thop library. Upon the integration of the LFA module to generate local salient features, the model does not introduce any additional parameters during the feature map segmentation stage. However, in the local facial expression feature enhancement stage, four sets of convolutional layers are incorporated. Each local feature map with c channel numbers uses two convolutional layers with the convolutional kernel size of k for feature enhancement. At this time, the spatial complexity of the LFA module is $o(8c^2k^2)$. Figure 14 illustrates the specific data of the overall parameter amount of the model. When adding LFA, its size becomes 21.58 M, and the parameter amount of the LFA module reaches 0.29 M. When the MGC module is added, the feature map of size (c, h, w) first needs to be extracted by grouping convolution in the block. The spatial complexity of the MGC module in this case is $O(c^2/4 + 3c^2k^2/2)$, where k represents the convolution kernel size. When the fusion convolutional self-attention mechanism is introduced, the spatial complexity of the MGC module reaches $O(3c^2 + 3k^2N + ck^2)$, where N represents the number of heads of the multi-head self-attention mechanism. The MGC module introduces only 0.07 M throughout the model.

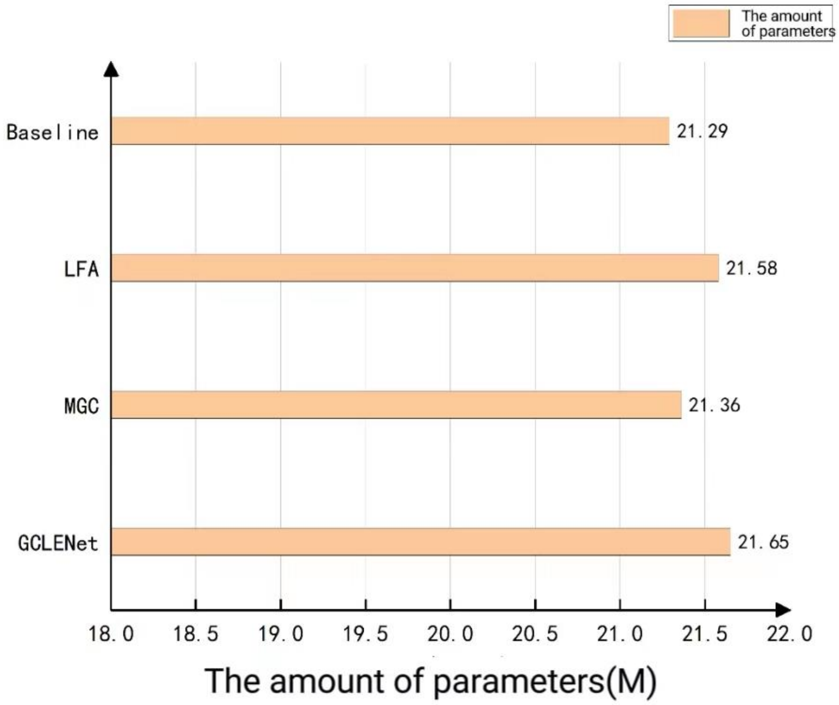


Figure 14. The number of model parameters.

The evaluation of floating-point operations per second (FLOPs) for each model is shown in Figure 15. The FLOPs of the baseline model ResNet are 3678.23 M. After the LFA module is added, the feature map segmentation consumes time to process, and the introduced four sets of convolution calculations occupy all the overhead of the LFA module. The time complexity of the LFA module is $O(2hwc^2k^2)$. Since the LFA module is located in the underlying structure of the entire model, the actual calculation amount of the model reaches 3911.05 M. After the MGC module is added, the model operation amount is almost the same as in the previous case. The time complexity of multi-scale convolution is $O(hwc^2/4 + 3hwc^2k^2/2)$, and the time complexity of the introduced fusion convolution self-attention mechanism is $O(3hwc^2 + chw(4k_c^2 + 2k_a^2 + k_c^4))$, where k_c represents the convolution kernel size of convolution computation and k_a represents the convolution kernel size of the self-attention mechanism. Thus, the FLOPs of the model after adding MGC reach 3905.07 M.

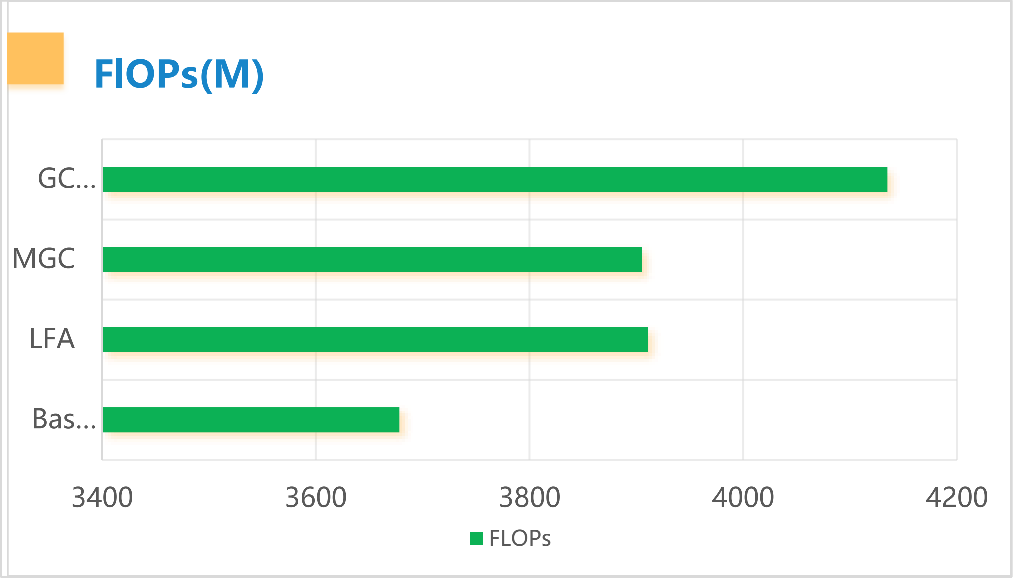


Figure 15. Comparison of FLOPs for the models.

In this chapter, we conducted thorough experimental validation and scientific discussions of the performance of the LSDC-FER algorithm. The experimental design took into full consideration the diversity and representativeness of datasets, selecting standard facial expression datasets such as RAF-DB and CK+. We meticulously chose modern algorithms as comparison baselines to ensure the comprehensiveness and objectivity of our experimental results. The results indicated a significant improvement in recognition accuracy, particularly under complex scenarios, where the proposed algorithm demonstrated remarkable robustness in handling occlusion and pose variations. Ablation experiments validated the crucial role of the global information association module and local feature enhancement module in improving algorithm performance. The results of the study confirmed the importance of the organic integration of multi-scale feature extraction with the self-attention mechanism, as well as the effective fusion of local and global information, as key avenues for improving FER algorithm performance.

5. Conclusions

In this paper, we present a novel approach for face expression recognition, which utilizes local features to enhance the integration of global information. Our method incorporates a local feature enhancement module to enhance the extraction of local features. This module effectively directs the model's attention to specific local features by segmenting the feature map. Additionally, asymmetric convolution techniques are employed to further enhance the extraction of local features. This results in significant improvements in local feature representation. Furthermore, we incorporate a global information association module in our approach. This module utilizes diverse convolution kernels to extract multi-scale information from facial expressions. Moreover, we combine fusion convolution with self-attention mechanisms to effectively extract and consolidate the associated information among expression features. This enables a comprehensive understanding of the context and relationships within facial expressions. Lastly, we combine the extracted local features, global multi-scale features, and global context information to mitigate the interference caused by occlusion and pose transformation. By combining these diverse types of features, our aim is to enhance the model's robustness and ensure its effectiveness in addressing variations caused by occlusion and pose transformations. Experimental results demonstrate that our method significantly improves the performance of face expression recognition in natural environments. It exhibits enhanced robustness and generalization capabilities in comparison with the existing approaches.

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

Funding: This work is funded by the National Natural Science Foundation of China (61772180), the Hubei Provincial Natural Science Foundation (2023BCB041), and the Hubei Provincial Education Science Planning Project (2022GB030).

Data Availability Statement: The data that support the findings of this study are available from the corresponding author (Lingyu Yan), upon reasonable request.

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

References

1. Yan, L.; Sheng, M.; Wang, C.; Gao, R.; Yu, H. Hybrid neural networks based facial expression recognition for smart city. *Multimed. Tools Appl.* **2022**, *81*, 319–342. [CrossRef]
2. Corneanu, C.A.; Simón, M.O.; Cohn, J.F.; Guerrero, S.E. Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1548–1568. [CrossRef] [PubMed]
3. Vithanawasam, T.M.W.; Madhusanka, B. Dynamic face and upper-body emotion recognition for service robots. In Proceedings of the 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), Singapore, 6–8 June 2018; pp. 428–432.
4. Kapoor, A.; Burleson, W.; Picard, R.W. Automatic prediction of frustration. *Int. J. Hum.-Comput. Stud.* **2007**, *65*, 724–736. [CrossRef]
5. Irani, R.; Nasrollahi, K.; Simon, M.O.; Corneanu, C.A.; Escalera, S.; Bahnsen, C.; Lundtoft, D.H.; Moeslund, T.B.; Pedersen, T.L.; Klitgaard, M.-L.; et al. Spatiotemporal analysis of RGB-DT facial images for multimodal pain level recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 88–95.
6. Yang, H.; Liu, L.; Min, W.; Yang, X.; Xiong, X. Driver yawning detection based on subtle facial action recognition. *IEEE Trans. Multimed.* **2020**, *23*, 572–583. [CrossRef]
7. Darwin, C.; Prodger, P. *The Expression of the Emotions in Man and Animals*; Oxford University Press: New York, NY, USA, 1998.
8. Ekman, P.; Friesen, W.V. Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol.* **1971**, *17*, 124–129. [CrossRef] [PubMed]
9. Ekman, P.; Friesen, W.V. *Facial Action Coding System: Investigator's Guide*; Consulting Psychologists Press: Palo Alto, CA, USA, 1978.
10. Platt, J. *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*; Microsoft Research: Mountain View, CA, USA, 1998.
11. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [CrossRef]
12. MacQueen, J. Classification and analysis of multivariate observations. In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 21 June–18 July 1965; University of California, Berkeley: Berkeley, CA, USA, 1966; pp. 281–297.
13. Cootes, T.F.; Taylor, C.J.; Cooper, D.H.; Graham, J. Active shape models-their training and application. *Comput. Vis. Image Underst.* **1995**, *61*, 38–59. [CrossRef]
14. Matthews, I.; Baker, S. Active appearance models revisited. *Int. J. Comput. Vis.* **2004**, *60*, 135–164. [CrossRef]
15. Setyati, E.; Suprapto, Y.K.; Purnomo, M.H. Facial emotional expressions recognition based on active shape model and radial basis function network. In Proceedings of the 2012 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications (CIMSA) Proceedings, Tianjin, China, 2–4 July 2012; pp. 41–46.
16. Han, S.; Meng, Z.; Liu, P.; Tong, Y. Facial grid transformation: A novel face registration approach for improving facial action unit recognition. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 1415–1419.
17. Bashyal, S.; Venayagamoorthy, G.K. Recognition of facial expressions using Gabor wavelets and learning vector quantization. *Eng. Appl. Artif. Intell.* **2008**, *21*, 1056–1064. [CrossRef]
18. Zhang, B.; Liu, G.; Xie, G. Facial expression recognition using LBP and LPQ based on Gabor wavelet transform. In Proceedings of the 2016 2nd IEEE International Conference on Computer and Communications (ICCC), Chengdu, China, 14–17 October 2016; pp. 365–369.
19. Yang, H.; Ciftci, U.; Yin, L. Facial expression recognition by de-expression residue learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2168–2177.

20. Ruan, D.; Yan, Y.; Lai, S.; Chai, Z.; Shen, C.; Wang, H. Feature decomposition and reconstruction learning for effective facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7660–7669.
21. Zhao, Z.; Liu, Q.; Wang, S. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Trans. Image Process.* **2021**, *30*, 6544–6556. [CrossRef]
22. 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]
23. 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.
24. Zhang, Y.; Wang, C.; Deng, W. Relative uncertainty learning for facial expression recognition. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 17616–17627.
25. Li, S.; Deng, W.; Du, J.P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2852–2861.
26. 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.
27. Deng, J.; Pang, G.; Zhang, Z.; Pang, Z.; Yang, H.; Yang, G. cGAN based facial expression recognition for human-robot interaction. *IEEE Access* **2019**, *7*, 9848–9859. [CrossRef]
28. Yovel, G.; Duchaine, B. Specialized face perception mechanisms extract both part and spacing information: Evidence from developmental prosopagnosia. *J. Cogn. Neurosci.* **2006**, *18*, 580–593. [CrossRef] [PubMed]
29. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
30. Cai, J.; Meng, Z.; Khan, A.S.; Li, Z.; O'Reilly, J.; Tong, Y. Probabilistic attribute tree in convolutional neural networks for facial expression recognition. *arXiv* **2018**, arXiv:1812.07067. [CrossRef]
31. Xie, S.; Hu, H. Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks. *IEEE Trans. Multimed.* **2018**, *21*, 211–220. [CrossRef]
32. 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.
33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
35. 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.
36. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
37. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
38. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 1597–1607.
39. Dapogny, A.; Bailly, K.; Dubuisson, S. Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection. *Int. J. Comput. Vis.* **2018**, *126*, 255–271. [CrossRef]
40. Li, Y.; Zeng, J.; Shan, S.; Chen, X. Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Trans. Image Process.* **2018**, *28*, 2439–2450. [CrossRef]
41. Pan, X.; Ge, C.; Lu, R.; Song, S.; Chen, G.; Huang, Z.; Huang, G. On the integration of self-attention and convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 815–825.
42. Li, Y.; Zeng, J.; Shan, S.; Chen, X. Patch-gated CNN for occlusion-aware facial expression recognition. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 2209–2214.
43. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **2017**, *10*, 18–31. [CrossRef]
44. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
45. Cai, J.; Meng, Z.; Khan, A.S.; O'Reilly, J.; Tong, Y. Island loss for learning discriminative features in facial expression recognition. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 302–309.
46. Xie, S.; Hu, H.; Wu, Y. Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. *Pattern Recognit.* **2019**, *92*, 177–191. [CrossRef]

47. Fan, Y.; Lam, J.C.K.; Li, V.O.K. Multi-region ensemble convolutional neural network for facial expression recognition. In *Artificial Neural Networks and Machine Learning—ICANN 2018: 27th International Conference on Artificial Neural Networks. Rhodes, Greece, 4–7 October 2018*; Proceedings, Part I 27; Springer International Publishing: Cham, Switzerland, 2018; pp. 84–94.
48. Lee, J.; Kim, S.; Kim, S.; Park, J.; Sohn, K. Context-aware emotion recognition networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10143–10152.
49. Zhao, G.; Huang, X.; Taini, M.; Li, S.Z.; PietikäInen, M. Facial expression recognition from near-infrared videos. *Image Vis. Comput.* **2011**, *29*, 607–619. [CrossRef]
50. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
51. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
52. Kuo, C.M.; Lai, S.H.; Sarkis, M. A compact deep learning model for robust facial expression recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2121–2129.

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.