

Article

Facial Expression Recognition Based on Vision Transformer with Hybrid Local Attention

Yuan Tian , Jingxuan Zhu, Huang Yao and Di Chen * 

Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan 430079, China; tianyuan_ccnu@163.com (Y.T.); zjx_ccnu@163.com (J.Z.); yaohuang@ccnu.edu.cn (H.Y.)

* Correspondence: chendi@ccnu.edu.cn

Abstract: Facial expression recognition has wide application prospects in many occasions. Due to the complexity and variability of facial expressions, facial expression recognition has become a very challenging research topic. This paper proposes a Vision Transformer expression recognition method based on hybrid local attention (HLA-ViT). The network adopts a dual-stream structure. One stream extracts the hybrid local features and the other stream extracts the global contextual features. These two streams constitute a global–local fusion attention. The hybrid local attention module is proposed to enhance the network’s robustness to face occlusion and head pose variations. The convolutional neural network is combined with the hybrid local attention module to obtain feature maps with local prominent information. Robust features are then captured by the ViT from the global perspective of the visual sequence context. Finally, the decision-level fusion mechanism fuses the expression features with local prominent information, adding complementary information to enhance the network’s recognition performance and robustness against interference factors such as occlusion and head posture changes in natural scenes. Extensive experiments demonstrate that our HLA-ViT network achieves an excellent performance with 90.45% on RAF-DB, 90.13% on FERPlus, and 65.07% on AffectNet.

Keywords: facial expression recognition; attention; vision transformer



Citation: Tian, Y.; Zhu, J.; Yao, H.; Chen, D. Facial Expression Recognition Based on Vision Transformer with Hybrid Local Attention. *Appl. Sci.* **2024**, *14*, 6471. <https://doi.org/10.3390/app14156471>

Academic Editor: Thomas Lindner

Received: 25 June 2024

Revised: 21 July 2024

Accepted: 22 July 2024

Published: 24 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

In recent years, with the rapid development of computer technology and artificial intelligence, the field psychologist Mehrabian [1] pointed out, through a large number of experimental studies, that only 7% of the emotional information exchanged in people’s daily communication comes from verbal expressions. Voice intonation accompanying verbal expressions accounts for 38%, while approximately 55% of the emotional communication information is conveyed through people’s facial expressions. With the development of computer technology and artificial intelligence, the computer recognition of facial expressions has become a reality. A facial expression recognition system outputs the expression category corresponding to the face in the image for a given face image, and these research results can be used in various fields such as human–robot interaction, safe driving, medical care, and education.

In the field of human–robot interaction, the robots obtain emotional information from human expressions, making the interaction process more intelligent and enabling robots to have human empathy and, thus, respond better; in the field of safe driving, the facial expression recognition system monitors the driver’s mental state, determines whether the driver’s driving state is safe or not, and issues alerts to remind the driver in time. In the medical field, the facial expression recognition system can analyze whether patients have pain or other abnormal states, and assist in treatment. In the education field, the facial expression recognition system can monitor students’ expressions in distance education or regular education, and obtain students’ mental state when learning, and give feedback

in a timely manner to assist in teaching. From the above application cases, it is clear that facial expression recognition has a wide application prospect and has important research significance and application value.

With the rapid development of related computing devices and algorithms, significant progress has been made in the field of face expression recognition. The feature extraction methods have gradually evolved from the early manual feature extraction to adaptive feature extraction using deep-learning methods. The performance of face expression recognition has improved significantly in various published datasets, and the performance has reached saturation on some datasets under traditional controlled conditions like CK+ [2]. However, facial expression recognition in natural scenes faces many challenges. The dataset in the natural scene is different from the traditional control condition dataset, such as the source of the face image being different, and the face expression image in the dataset is obtained through the Internet, TV series, movies, etc. Under controlled conditions, expressions are generated through guidance, while expressions in natural scenes are spontaneous. In addition, the main challenges of facial recognition tasks in natural scenes include face occlusions and head pose variation. Face occlusion results in the loss of face information in the occluded part, which is also a common problem affecting the accuracy of expression recognition in natural scenes. Head pose variation will change the shape of the face, blur the facial expression information, make the expression information distributed in different positions and angles, and affect the correct recognition.

With the advancement in deep learning, most expression recognition methods use convolutional neural networks (CNNs) for extracting image features. These methods have demonstrated exceptional performance in the field of expression recognition. However, CNNs are implemented by continuously overlapping convolutional layers to perform image extraction from local to global, which not only leads to a significant increase in computation, but is also accompanied by the problem of gradient disappearance, making the network unable to converge. In addition, these CNN-based expression recognition methods also lack robustness and can be affected by occlusion and head pose variation, failing to focus on more important facial regions. To address the above problems, this paper proposes a Vision Transformer expression recognition method based on hybrid local attention. The contributions are as follows:

(1) We propose a novel hybrid local attention module to address the challenge of expression recognition in natural scenes and enhance the overall performance of expression recognition.

(2) The HLA-ViT network utilizes a dual-stream structure, where one stream focuses on extracting hybrid local features and the other stream focuses on capturing global contextual features. These streams work together to form a comprehensive global-local fusion attention mechanism. Furthermore, we employ a decision-level fusion strategy to effectively combine the classification results from these two streams.

(3) Through extensive experimentation, we demonstrate the efficiency and effectiveness of our proposed HLA-ViT approach. Our method achieves an excellent performance on three widely used datasets, RAF-DB, FERPlus, and AffectNet.

2. Related Work

The most common network in deep-learning-based facial expression recognition methods is the convolutional neural network [3]. The VGG-16 and VGG-19, for example, are CNN models proposed by Simonyan and Zisserman [4], whose name comes from the acronym of the Visual Geometry Group (VGG) at the University of Oxford, where the authors work. The model participated in the ImageNet Image Classification and Localization Challenge in 2014 and achieved excellent results. It ranked second in the classification task and first in the localization task. The VGG network is characterized by a simple structure and deeper layers. It uses the same convolutional kernel parameters in all the convolutional layers, and the model is constructed by stacking several convolutional and pooling layers. He et al. proposed the ResNet-18, ResNet-34, ResNet-50, and ResNet-152

networks [5], and their models achieved five firsts on ImageNet. ResNet modifies VGG19 by adding a residual unit through a short-circuiting mechanism so that the degradation of the network is well-solved. DenseNet-161 [6] connects two arbitrary layers of the same feature map size. This makes it easy to design networks with hundreds of layers and eliminates optimization difficulties. In the experiments, the accuracy of DenseNet increased with the number of parameters and there was no poor performance or overfitting. This structure learns simpler and more accurate models by reusing network features. Akhand et al. [7] proposed a deep convolution neural network based on transfer-learning technology. This method addressed the problem where most existing facial expression recognition methods consider only frontal images and ignore side images for simplicity of training. The proposed facial expression recognition system used eight different pre-trained DCNN models on the KDEF and JAFFE datasets for experimental validation, and the final pre-trained models achieved a high accuracy on both datasets. It has important reference value for the development of the subsequent expression recognition system. Howard et al. [8] used deep separable convolution to build lightweight deep neural networks and proposed a series of network architectures called MobileNets. The superiority of MobileNets was demonstrated through experiments on tasks such as object detection, fine-grained classification, and face detection. Sadik et al. [9] applied the MobileNet model to facial expression analysis by transfer learning. The experimental results showed that the MobileNet model with the transfer-learning approach can provide satisfactory results in recognition tasks. Agrawal and Mittal [10] proposed two novel convolutional neural network architectures and investigated the effect of the size of the convolutional kernel and the number of filters on the classification accuracy in CNNs. The impact of these two indicators on network accuracy was verified through experiments on the FER2013 dataset.

The dual-stream convolutional neural network is often used to enhance the feature extraction capability. Simmoyan [11] proposed a groundbreaking dual-stream convolutional neural network structure by integrating spatial and temporal motion information in video as the input of the two networks, respectively. The dual-stream network combines the effective information in both networks, and classifies still images using a deep convolutional network. Then, a temporal deep convolutional neural network is used to train and classify the optical flow information in the video stream. Finally, Softmax is used to fuse the two streams, so that the output of the network is better than that of the single stream network. Jung [12] proposed a dual-stream network to better identify useful features of facial expressions. Its deep network is based on two streams, where the first network extracts temporal features from image sequences and the other network extracts geometric features from facial marker points, which are finally fused at the decision layer and output by Softmax. The method achieves excellent performance in the CK+ dataset. Feichtenhofer [13] proposed a new convolutional network architecture for the spatio-temporal fusion of video clips. The two neural networks are fused at one layer. Extensive experiments showed that the fusion of spatio-temporal networks in the convolutional layer will not only improve the recognition performance of the network, but also reduce the parameters of the network. Feichtenhofer et al. [14] also proposed a video action recognition method based on spatio-temporal residual networks. The pre-trained ResNets was used to initialize the model, and the pooling layer was used between the two streams.

In order to better exploit the advantages of deep learning for facial expression recognition, many researchers have improved the convolutional neural network to improve the feature representation ability without changing the architecture of the network model itself. The commonly used methods include introducing the attention mechanism or using an improved activation function or loss function.

The attention mechanism is crucial in focusing on contextually relevant features in input feature maps in image classification [15]. By allowing neural networks to ignore irrelevant information, the attention mechanism helps networks to focus on valid information. Recently, Vision Transformer [16], which is based on the self-attention mechanism, has achieved very good results in the field of image processing. The Vision Transformer

model has decomposed images into image blocks and used linear embedded sequences of these blocks as input to the Transformer to perform image classification tasks. A spatial transformer module is proposed through an attention mechanism to transform the spatial domain information of an image to another corresponding space to extract the region of interest in the image [17]. Hu et al. [18] proposed the Squeeze-and-Excitation (SE) block. It processes the channel features in the input data during the training process. However, the SE block only considers the role of the channel attention mechanism in the training process, ignoring the spatial domain information in the image data. The spatial domain transformation network also has a similar problem, ignoring the influence of channel attention. Woo et al. [19] proposed the Convolutional Block Attention Module (CBAM), which is an attention mechanism module that effectively combines spatial attention and channel attention, unlike the spatial transformer module and SE block. Le et al. [20] proposed a new deep network that efficiently recognizes human emotions using a new global attention mechanism. Cao et al. [21] proposed a method to address the lack of feature extraction capability of the network by combining CBAM with VGGNet.

In this paper, we propose a local hybrid attention mechanism consisting of improved channel attention and spatial attention. It enables the network to focus on locally salient information with a better robustness to facial occlusion and head pose changes.

3. Methodology

The overall architecture of the proposed HLA-ViT is shown in Figure 1. It introduces a hybrid attention mechanism to solve the problem of expression recognition in natural scenes and improve the expression recognition performance of the model. Specifically, HLA-ViT transforms the feature maps extracted by the CNN as a backbone network into visual sequences by using ViT, and learns the contextual connections between these visual blocks using a global self-attention mechanism. The global self-attention mechanism enables the model to guide the network to learn facial features with robust occlusion and head pose from a global perspective under the condition of missing information, and the network has strong robustness. Meanwhile, the local attention module is introduced in the feature extraction backbone of HLA-ViT, which enables the network to focus on locally salient information, and the facial expression features are complemented to eliminate the interference of useless information. Finally, a decision-level fusion mechanism is used to fuse the expression information of the two parts to obtain the classification results and improve the expression recognition performance of the network in natural scenes.

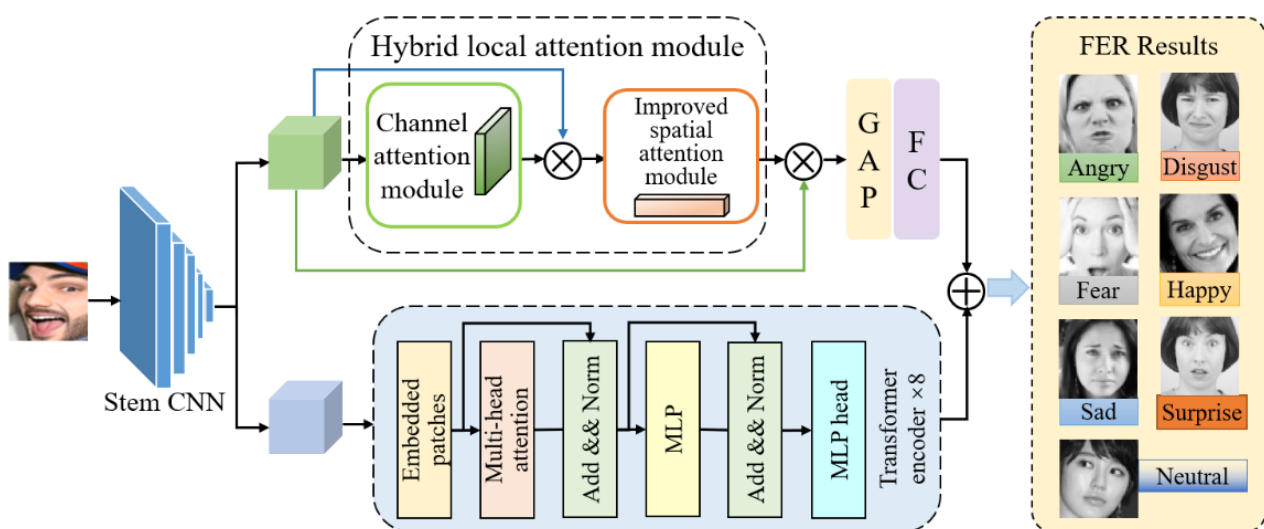


Figure 1. The proposed HLA-ViT network architecture.

HLA-ViT mainly includes two modules, the hybrid local attention module and the multi-layer Transformer encoders. The model is built on an IR50 network loaded with

pre-trained parameters, using the first three stages as the backbone network to extract the features from the low layers of the image. Then, a two-stream network is designed to process the extracted feature maps to obtain the global features and local features.

One stream uses a hybrid local attention module to combine the input features, enhance local information through spatial attention and channel attention, and learn local salient features. Then, the global average pooling layer and the fully connected layer are taken for classification. The other stream extends the obtained feature maps into a one-dimensional image feature sequence, learns the global contextual relationships of the feature sequences, and uses multi-layer Transformer encoder for expression classification. Finally, a decision-level fusion mechanism is employed to integrate local features with global feature classification results, further enhancing complementary information. This enables the model to dynamically update prominent features in the feature map, allowing the network to learn the most representative features among multiple characteristics, thereby improving the recognition performance of the network.

3.1. Convolutional Feature Extraction Network

CNNs are often used as the backbone for extracting image features; ResNet [5] is proposed to solve the problem of gradient dispersion caused by the deepening of the network. The residual unit proposed by ResNet makes it more efficient than normal networks by adding a “short-circuiting” mechanism between each of the two layers to obtain the feature parts without redundancy, and then using nonlinear activation for the redundant information to form the residual learning. The backbone network used in this paper is Improved ResNet-50 (IR50) [22]. IR50 optimizes the problems that exist in ResNet. For the residual module, the ReLU activation function and the last BN layer on the main propagation path are moved to the beginning of the Residual Block, and the BN layer is added at the end of the main propagation path. For the projection shortcut module, the maximum pooling is added to the beginning of the 1×1 convolution to avoid the information loss problem caused by the dimensionality reduction. In this paper, the input image size is $3 \times 224 \times 224$, then downsampled to $3 \times 112 \times 112$, and then fed into the model.

IR50 has four convolution stages to extract features, similar to ResNet, but they differ in the number of output feature maps. In this paper, only the first three stages of IR50 are utilized to extract the feature maps, and the size of the extracted feature maps is $256 \times 14 \times 14$, ensuring that the subsequent input to the Vision Transformer meets the conditions for patch embedding.

3.2. Hybrid Local Attention Module

The hybrid local attention module proposed in this paper consists of a combination of spatial attention and channel attention. When the dimensionality of the input features is $H \times W \times C$, the new feature dimensions obtained remain the same after passing the hybrid attention module. The previous method used to segment faces into blocks through facial landmarks or random cropping to obtain effective local features, thereby eliminating interference from facial occlusion and pose variations. However, these methods may lead to misalignment or uncertainty in facial expression recognition. In this paper, the feature mapping based on spatial domain and channel domain information is combined, and the feature mapping is passed through the hybrid attention module, so as to obtain the feature mapping with enhanced important local features. The network first learns what features are key features through the channel attention module, and then uses the spatial attention module to learn where the key features are, enhancing the acquisition of discriminative features in images. The hybrid local attention module is shown in Figure 2.

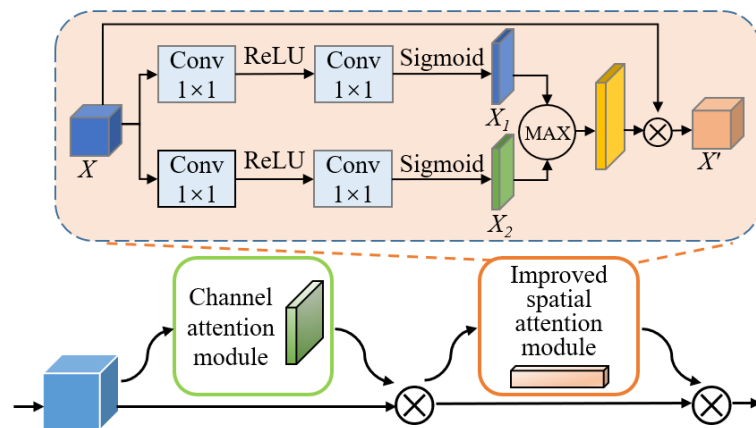


Figure 2. Hybrid local attention module.

The spatial attention module provides varying degrees of attention to different regions of the input image or feature image, increasing the influence of the target region and reducing the influence of irrelevant regions. The spatial attention in CBAM proposed by Woo [19] is to merge two $H \times W \times 1$ feature maps obtained by maximum pooling and average pooling to obtain an $H \times W \times 2$ feature map. Then, a spatial attention weight matrix of $H \times W \times 1$ is output by convolution and sigmoid function. In this paper, the proposed spatial attention module replaces the maximum pooling and average pooling operations with a two-layer 1×1 convolutional layer with ReLU and sigmoid activation functions. Let $X \in R^{H \times W \times C}$ denote the input feature maps, where H , W , and C represent the height, width, and number of feature maps, respectively. After the first 1×1 convolutional layer, output C/r feature maps, where r represents the reduction rate of the feature maps. The ReLU layer is then used to enhance the nonlinear relationship between the layers of the network. After the second 1×1 convolution layer, the number of feature maps is reduced to 1, and the attention maps are generated by the Sigmoid function. Thus, through two successive 1×1 convolution layers, the model automatically locates the important facial parts.

Meanwhile, in order to further enhance spatial attention module, a local attention network, which is a two-branch spatial attention module, is proposed. It is similar to the operation in CBAM, generating two spatial attention maps X_1 and X_2 , but, instead of combining them for the convolution operation, the final attention map X_{out} is generated by taking the maximum value of each attention map element in the two spatial attention maps. This can guide the model to focus on feature regions that are more useful for facial expression recognition. The specific formula is shown as follows:

$$X_{out}(w, h) = \text{MAX}\{X_1(w, h), X_2(w, h)\} \quad 1 \leq w \leq W, 1 \leq h \leq H \quad (1)$$

In Equation (1), H and W represent the height and width of the feature map, respectively. Finally, X_{out} is multiplied with the original feature map X element by element, so that the unimportant regions in the original feature map are suppressed, and the role of important regions in expression recognition is highlighted.

The specific structure of the local attention network is an improvement module of spatial attention. When a feature map X with dimension $H \times W \times C$ is the input to the local attention module, the computational formula to obtain the new feature map X' is shown as follows:

$$X_1 = \text{Sigmoid}(\text{Conv}(\text{ReLU}(\text{Conv}(X)))) \quad (2)$$

$$X_2 = \text{Sigmoid}(\text{Conv}(\text{ReLU}(\text{Conv}(X)))) \quad (3)$$

$$X_{out} = \text{MAX}(X_1, X_2) \quad (4)$$

$$X' = X_{out} \otimes X \quad (5)$$

R represents the reduction rate of the number of channels, and the attention matrix of $H \times W \times 1$ can be multiplied by the initial feature map to obtain the enhanced spatial attention feature map.

For the channel attention module, the approach used in this paper is also different from CBAM. In this paper, ECA-Net [23] is used as the channel attention module. It can effectively explain the importance of each channel in feature extraction. At the same time, its attention mechanism can adaptively adjust according to the input feature map, which can better adapt to feature extraction tasks in different scenarios.

In the hybrid local attention module proposed in this paper, the input feature map x is successively passed through the improved spatial attention module and the channel attention module, and then obtains the hybrid local attention feature image y with the same size as the input feature map. Each attention module has a dot product operation between the attention matrix and the original image, so that the feature map will go to the next module with the enhanced attention. This hybrid local attention module does not change the dimensionality of the feature maps and is a plug-and-play attention that can be directly integrated into the IR50 backbone network discussed previously. The network after the hybrid attention module can obtain the significant local information of the expression images more effectively, which is a beneficial contribution to the subsequent classification of expression recognition.

3.3. Vision Transformer with Hybrid Local Attention

Vision Transformer learns the features of images through a global self-attention mechanism. The application of Vision Transformer to facial expression recognition can improve the network's robustness to occlusion and head pose variations in natural scenes. We combine Vision Transformer with the proposed hybrid local attention module for facial expression recognition, which can effectively extract local and global features of the image. Fusion strategies is adopted to further improve the model's expression ability.

The specific operation is to expand the feature map sequence x extracted from the backbone IR50 into a linear sequence by flattening operation and inputting it into the multilayer Transformer encoder. The dimension of the feature map sequence x is $256 \times 14 \times 14$, where 14×14 is the feature map space size and 256 is the channel size. The 2D feature map needs to be flattened into a 1D visual embedding sequence, so the model reshapes the feature map x into a flat sequence and feeds it into a linear projection to obtain a 196×768 feature vector. In ViT, it is also necessary to append classification tokens (cls) to the beginning of the input sequence of the final prediction. In addition, a one-dimensional learnable position embedding is added to the feature embedding to obtain the position information of the image blocks. The multi-layer Transformer encoder consists of a sequence of L attention modules, where the size of L in this paper is 8. Each sequence consists of a multi-headed self-attention (MHSA) module and a multilayer perceptron (MLP), where the number of self-attention heads is set to 12. The MLP consists of two fully connected feature projection layers and a GELU for nonlinearity. Finally, a classification head implemented by a single-layer MLP is used for the classification output.

Two traditional fusion methods are feature-level fusion and decision-level fusion when choosing a fusion strategy. Feature-level fusion directly combines the feature maps of two streams into a joint feature vector, and then trains the classifier, while decision-level fusion combines the classification results of these two streams. Safont et al. [24] proposed a separated score integration method based on alpha integrals, and experimental results show that this method performs better than the considered single classifier and classical fusion techniques. Salazar et al. [25] studied the decision-level fusion method and proposed two new graph-based regularization methods, and the experimental results proved the superiority of the method. In HLA-ViT, the hybrid local features extracted by the network and the global contextual features extracted by the multi-layer Transformer encoder are complementary at the feature level, constituting global-local fusion attention. Therefore, the network model in this paper adopts a decision-level fusion strategy.

Decision-level fusion is the highest level of fusion, which requires modeling emotional information separately for each channel, and then merging the recognition results of all channels. The advantage of decision-level fusion is its ability to parallelly combine multiple classifiers, independently integrating each classifier. This allows them to work autonomously, enabling each sub-classifier result to contribute to the final classification decision through a rational approach. When combining multiple classifiers in a parallel manner, the results of each sub-classifier can be in the form of classification probabilities, classification distances, actual classification results, or relevance metrics for different information classes. Typically, in practical expression recognition systems, they are designed as classification results. Because each sub-classifier works independently in a parallel combination, the output information of each classifier does not affect others.

The local and global features obtained through the previous method are, respectively, input into the corresponding branch networks for emotional category prediction. The loss functions used in both streams are cross-entropy loss functions, which can be expressed as follows:

$$Loss_i = -\frac{1}{N} \sum_{i=0}^{N-1} \log \frac{e^{W_{y_i}^{(k)T} v_i^{(k)} + b_{y_i}^{(k)}}}{\sum_{j=0}^{C-1} e^{W_j^{(k)T} v_i^{(k)} + b_j^{(k)}}} \quad (6)$$

where N is the number of samples, C is the number of categories for classification, $W^{(k)}$ and $b^{(k)}$ are the weight matrix and offset of the fully connected layer, respectively, $v_i^{(k)}$ is the i th input of the fully connected layer, y_i is the category label, and $i \in \{local, global\}$.

The final model loss function is shown as follows:

$$Loss = \lambda_1 Loss_{local} + \lambda_2 Loss_{global} \quad (7)$$

where λ_1 and λ_2 are hyperparameters to balance the two streams. λ_1 represents the weight proportion of the recognition rate obtained using the HLA module, while λ_2 represents the weight proportion of the recognition rate obtained using the ViT module. During the model training process, parameters are updated based on the loss values, and multiple iterative training steps are performed to minimize the loss function.

4. Experimental Results

4.1. Implementation Details

The input batch image data are cropped to a fixed size, 224×224 , and randomly flipped horizontally to avoid overfitting. At the same time, the data enhancement method RandomErasing is used to further improve the performance of the model. In this paper, the basic network IR50 pre-trained on the MS1MV3 [26] face recognition dataset is used as the feature extraction network. Face features are extracted from the third convolution stage of IR50. The depth of the Transformer encoders is set to 8 in this experiment, and the MLP ratio and drop path are set to 2 and 0.01, respectively. The training batch size is set to 128 and the learning rate is initialized to 4×10^{-5} using the ExponentialLR learning rate decay. The entire network is optimized using the Adam optimizer. Using the Adam optimizer allows the network to converge faster and makes tuning the parameters more convenient. Use the standard label smoothing cross-entropy loss to supervise the model, which has better generalization for expression recognition. The method is implemented using the Pytorch 1.11.0 toolbox, with the initial momentum coefficient of the network set to 0.9 and trained on a single NVIDIA RTX 3090 GPU (NVIDIA, Santa Clara, CA, USA).

The experiments are evaluated on three different commonly used expression datasets in a natural scene. The RAF-DB dataset contains samples of neutral expressions and six basic expression categories labeled with 12,271 images for training and 3068 images for testing. The FERPlus database contains 28,709 training images and 3555 test images, and, in addition to the neutral expressions and six basic expressions, the labels also contain the contempt category. The AffectNet dataset includes six basic expressions and neutral expressions. The training and test sets contain 283,901 and 3500 images, respectively.

The experimental analysis on the RAF-DB dataset of the depth of the Transformer encoder used in HLA-ViT is shown in Figure 3. The experiment shows that the best results are achieved when using ViT with a depth of 8.

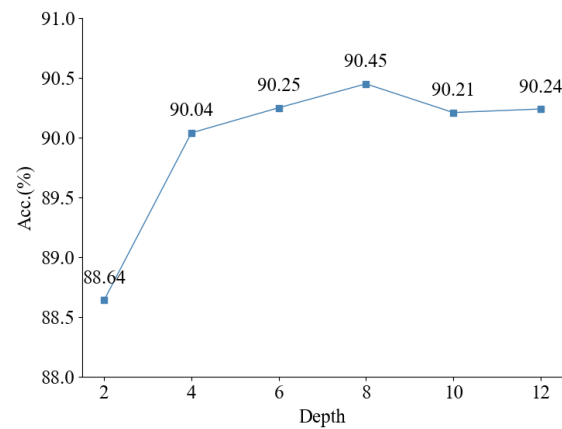


Figure 3. Evaluation of different numbers of transformer encoder on RAF-DB.

In order to explore the effect of the two hyperparameters λ_1 and λ_2 in the loss function on the HLA-ViT performance, the different recognition results on the RAF-DB dataset with different values from 0.1 to 0.9 are experimentally investigated. As shown in Table 1, when the two branches are fused at the decision level with a local loss weight of 0.4 and a global loss weight of 0.6, the facial expression recognition accuracy can reach 90.45%. This indicates that the importance of the ViT stream is slightly higher than that of the local attention stream. Subsequent ablation experiments will confirm that using only ViT or HLA for facial expression recognition does not achieve the same level of accuracy as when the two are fused. This indicates that the proposed method of fusing hybrid local features and global features can further enhance information complementarity and improve the recognition performance of the network.

Table 1. Evaluation of λ_1 and λ_2 on RAF-DB.

λ_1	λ_2	Acc. (%)
0.1	0.9	87.83
0.2	0.8	89.93
0.3	0.7	90.21
0.4	0.6	90.45
0.5	0.5	89.03
0.6	0.4	88.71
0.7	0.3	88.53
0.8	0.2	88.84
0.9	0.1	88.66

The results in Table 1 show that the importance of the ViT stream is slightly higher than that of the local attention stream for $\lambda_1 = 0.4$ and $\lambda_2 = 0.6$, when a good recognition performance is obtained.

4.2. Comparison with State-of-the-Art Methods

Results on RAF-DB: Table 2 evaluates HLA-ViT with the state-of-the-art methods on the RAF-DB dataset. We compare our method with the lately published approaches including SCN [27], PSR [28], RAN [29], DMUE [30], FDRL [31], FER-VT [32], Meta-Face2Exp [33], SPWFA-SE [34], and RCL-Net [35]. The experimental results demonstrate that the method proposed in this paper achieved a recognition rate of 90.45%, which is 0.98% better than the second-best method (FDRL [29]). Table 3 shows the class-wise accuracy on RAF-DB. As shown in Table 3, our method achieves the best results in most

categories including Angry, Disgust, Fear, Happy, and Neutral expressions. The RCL-Net [33] achieves higher accuracies in Sad (86%) and Surprise (88%), but its performances in Disgust (57%) and Fear (57%) are far from satisfactory.

Table 2. Comparison results with SOTA methods on RAF-DB, FERPLUS, and AffectNet.

Methods	Year	RAF-DB	FERPlus	AffectNet
SCN [27]	CVPR 2020	87.03	89.39	-
PSR [28]	CVPR 2020	88.98	-	63.77
RAN [29]	TIP 2020	86.90	89.16	-
DMUE [30]	CVPR 2021	89.42	-	63.11
FDRL [31]	CVPR 2021	89.47	-	-
FER-VT [32]	2021		90.04	
Meta-Face2Exp [33]	CVPR 2022	88.54	88.54	64.23
SPWFA-SE [34]	IEEE Trans 2023	86.31	-	59.23
RCL-Net [35]	2023	88.2	89.53	-
Ours	-	90.45	90.13	65.07

Table 3. Class-wise accuracy on RAF-DB, FERPLUS, and AffectNet datasets. Green, blue, and red mark the highest value of single category in RAF-DB, FERPLUS, and AffectNet, respectively.

Dataset	Method	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Mean Class Accuracy
RAF-DB	SPWFA-SE [34]	80	59	59	93	84	88	86	78
	RCL-Net [35]	86	57	57	95	86	88	89	79
	Ours	89	79	73	96	82	87	93	86
FERPlus	RCL-Net [35]	84	46	61	97	76	93	91	78
	Ours	86	56	62	96	81	89	91	80
AffectNet	Face2Exp [33]	57	56	64	87	66	56	62	64
	SPWFA-SE [34]	35	59	49	95	59	45	73	59
	Ours	57	61	60	88	61	61	67	65

Results on FERPlus: Table 2 displays the outcomes of a comparison of this paper's approach using the FERPLUS dataset with other state-of-the-art techniques. The FERPlus dataset is an enhanced version of the FER2013 dataset. Our HLA-ViT achieves the best accuracy of 90.13% on FERPlus. In Table 3, the proposed model showed a relatively good performance for Angry, Disgust, Fear, Sad, and Neutral expressions on the FERPlus dataset. The best performance was achieved in recognizing Happy expressions with a 96% accuracy, followed by Neutral, Surprised, Angry, and Sad with a 91%, 89%, 86%, and 81% accuracy, respectively, as well as Fear and Disgust with a 56% and 62% accuracy, respectively.

Results on AffectNet: The model presented in this paper is compared to several state-of-the-art methods in Table 2 for the AffectNet dataset. The images in AffectNet come from different backgrounds, races, and genders and include various poses, expressions, and lighting conditions. It focuses on real-life scenarios and captures human facial expression changes in different environments. Our method achieves the best accuracy of 65.07%, which is 0.84% higher than the second-best method (Meta-Face2Exp [33]). Table 3 illustrates the detailed comparison results on AffectNet. It can be seen that our proposed method achieves a superior performance in terms of the mean class accuracy. This indicates that our model has a better average recognition effect on all facial expression categories. Additionally, our method also achieves the highest accuracies for the three facial expressions (Angry, Disgust, and Surprise) among these methods, which are 57%, 61%, and 61%, respectively.

4.3. Ablation Study

Since HLA-ViT is composed of a hybrid local attention module (HLA) and ViT module after extracting features from the backbone, to verify the effectiveness of the modules, abla-

tion experiments are compared for each module on the RAF-DB, FERPlus, and AffectNet datasets, as shown in Table 4.

Table 4. Ablation study.

Method	HLA	ViT	RAF-DB	FERPlus	AffectNet
Baseline	×	×	87.22%	87.14%	61.02%
b	×	✓	90.16%	88.14%	63.84%
c	✓	×	89.93%	88.23%	62.34%
HLA-ViT	✓	✓	90.45%	90.13%	65.07%

IR50 is used as the baseline directly for the expression recognition task after pre-training. We individually investigated the impact of introducing the ViT module, the HLA module, and both on the model. These three methods correspond to method b, method c, and HLA-ViT in Table 4, respectively. The accuracy of the model on RAF-DB, FERPlus, and AffectNet is improved by 2.94%, 1.0%, and 2.82%, respectively, when method b is compared with Baseline, which indicates that, after applying ViT, the model obtains the global self-attention provided by ViT and enhances the network's ability to extract image features. The ViT has a strong robustness to the interference factors present in expression images in natural scenes. Method c is compared with Baseline, and the accuracy of the model on RAF-DB, FERPlus, and AffectNet is improved by 2.71%, 1.09%, and 1.32%, respectively, indicating the effectiveness of the HLA module, which also indicates that extracting local attention features is beneficial to improving the expression recognition performance. This can be explained by the fact that the local attention feature module enhances the robustness of the network to local occlusions and head pose changes, thus improving the performance of expression recognition in natural scenes. HLA-ViT combining hybrid local attention and global attention from ViT has improved the accuracy on RAF-DB, FERPlus, and AffectNet compared with other methods, which shows the effectiveness of both modules in this method.

4.4. Visual Analysis

Visualization techniques in experiments are also essential, and attention visualization enables a clearer intuitive view of the effectiveness of the local attention module and the global self-attention in ViT.

Figure 4 shows the attention maps under occlusion conditions. The first row is the original input images, the second row is the visualized images of the baseline method, and the third row is the visualized images of the proposed method HLA-ViT in this paper. The input images are the facial images selected from the dataset with occlusion. It can be observed that the utilization of the local attention module guides the network to focus attention on locally significant areas. These regions are crucial for the accurate recognition of facial expression categories. Compared with the baseline, the method proposed in this article can better focus on areas useful for facial expression recognition, such as eyes, mouth, and nose, under occlusion. For example, in the first column, where the mouth of the face in the original image is occluded, the baseline method concentrates attention on the nose and cheek regions while ignoring the eye regions. In contrast, our method directs attention to the eyes and nose, which are crucial for accurate expression recognition. Similarly, in the fourth column, where a person's hand covers a significant portion of the right side of the face in the original image, the baseline method fails to capture effective attention regions. However, our method adeptly focuses on the uncovered side of the face, specifically the eyes, nose, and mouth regions. The attention maps in the fifth and seventh columns also illustrate this issue. Results from the second, third, and sixth columns reveal that the baseline method tends to treat the occluded regions as interesting regions in cases of partial facial occlusion, leading to the introduction of useless information and, ultimately, obtaining incorrect recognition results. The combination of the hybrid local attention module and global attention in HLA-ViT enables the network to focus on unoccluded areas,

aligning with human perception. This further demonstrates the robustness of HLA-ViT in the presence of facial occlusion.

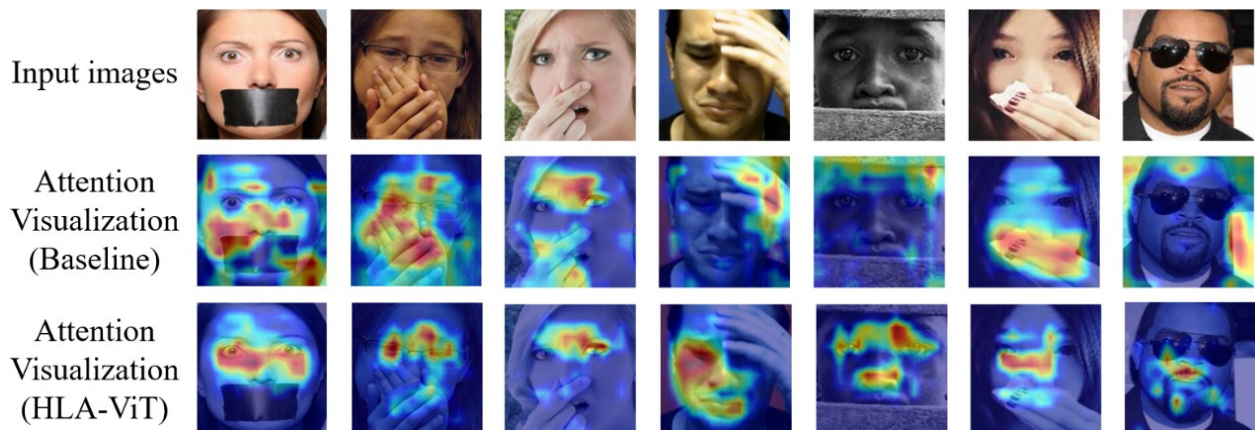


Figure 4. Attention visualization in the presence of occlusion.

Figure 5 focuses on the face images with variations in head pose, specifically selecting images with a lateral tilt of approximately 45° . The first row presents the original images, the second row displays results from the baseline method, and the third row shows the attention maps obtained through visualization using the proposed HLA-ViT method. From the images, it can be observed that, in the presence of variations in head pose, the baseline method sometimes misses the regions of interest that require attention, as shown in the first, third, and fourth columns. It may also overlook crucial areas for facial expression recognition, as seen in the sixth column where the mouth and eyebrows are essential for recognizing expressions. The baseline method neglects these areas, whereas our proposed approach utilizes the hybrid local attention module to concentrate more attention on the mouth and eyebrow regions. By leveraging the information conveyed by these focused areas, the network is better equipped to recognize and classify facial expressions.



Figure 5. Attention visualization in the presence of head variation.

Compared with the original image, it can be seen that the use of the local attention module can guide the network to focus on locally significant regions that are crucial for the correct recognition of expression categories. In addition, for faces with occlusions, without the local attention module, the network might focus on the occluded parts and ignore the highlighted parts of expressions, while the hybrid local attention module and the combination of global attention in HLA-ViT enable the network to focus on the non-occluded areas, which is consistent with human perception and demonstrates the robustness of HLA-ViT in face occlusions. For images with variations in head pose, HLA-ViT with the addition of

the hybrid local attention module can also focus more on important areas of the face, such as the mouth, eyes, etc. With the expression information expressed by these focal regions, the network is able to better classify and recognize their expressions.

5. Conclusions

This paper proposes a Vision Transformer model based on hybrid local attention, and applies the model to the expression recognition task. The approach addresses the expression recognition problem by converting facial images into visual text sequences and recognizing from a global perspective, while introducing an improved local attention module to enhance the robustness of the network to local occlusions and head pose changes. The decision-level fusion mechanism is adopted in the fusion strategy, using the visual Transformer to capture the robust features of the visual sequence context from a global perspective, and then fusing the locally significant expression features obtained by the local attention module to effectively utilize the complementary information between them and enhance the expression recognition performance. The experimental results demonstrate that the HLA-ViT exceeds other state-of-the-art methods on three frequently used facial expression datasets, i.e., RAF-DB, FERPlus, and AffectNet. The experimental results also verify that the proposed attention module in the network can effectively recognize the important facial regions to improve network performance.

Author Contributions: Conceptualization, Y.T., J.Z., H.Y. and D.C.; methodology, Y.T., J.Z., H.Y. and D.C.; software, J.Z.; validation, Y.T. and J.Z.; investigation, Y.T., J.Z., H.Y. and D.C.; writing—original draft preparation, Y.T. and J.Z.; writing—review and editing, Y.T. and J.Z.; supervision, Y.T., H.Y. and D.C.; project administration, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the General Project for Education of National Social Science Fund, Study on the Mechanism of Emotional Engagement and its Intervention in Primary and Secondary School Teachers' online Training (Grant Number: BCA230278).

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

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

References

1. Mehrabian, A.; Ferris, S. Inference of attitudes from nonverbal communication in two channels. *J. Consult. Psychol.* **1967**, *31*, 248–252. [[CrossRef](#)] [[PubMed](#)]
2. Lucey, P.; Cohn, J.; 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.
3. Zhang, J.; Yan, B.; Du, X.; Guo, Q.; Hao, R.; Liu, J.; Liu, L.; Ni, G.; Weng, X.; Liu, Y. Motion magnification multi-feature relation network for facial microexpression recognition. *Complex Intell. Syst.* **2022**, *8*, 3363–3376. [[CrossRef](#)]
4. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
5. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
6. Huang, G.; Liu, Z.; Maaten, L.; Weinberger, K. Densely connected convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
7. Akhand, M.; Roy, S.; Siddique, N.; Kamal, M.; Shimamura, T. Facial emotion recognition using transfer learning in the deep CNN. *Electronics* **2021**, *10*, 1036. [[CrossRef](#)]
8. Howard, A.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
9. Sadik, R.; Anwar, S.; Reza, M. AutismNet: Recognition of autism spectrum disorder from facial expressions using mobilenet architecture. *Int. J.* **2021**, *10*, 327–334.
10. Agrawal, A.; Mittal, N. Using CNN for facial expression recognition: A study of the effects of kernel size and number of filters on accuracy. *Vis. Comput.* **2020**, *36*, 405–412. [[CrossRef](#)]

11. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS), Cambridge, MA, USA, 8–13 December 2014; pp. 568–576.
12. Jung, H.; Lee, S.; Yim, J.; Park, S.; Kim, J. Joint fine-tuning in deep neural networks for facial expression recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2983–2991.
13. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941.
14. Feichtenhofer, C.; Pinz, A.; Wildes, R. Spatiotemporal multiplier networks for video action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4768–4777.
15. Zaman, K.; Zhaoyun, S.; Shah, B.; Hussain, T.; Shah, S.; Ali, F.; Khan, U. A novel driver emotion recognition system based on deep ensemble classification. *Complex Intell. Syst.* **2023**, *9*, 6927–6952. [\[CrossRef\]](#)
16. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16×16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), Parsa, Torabian, 3–7 May 2021; pp. 1–22.
17. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. Advances in neural information processing systems. In Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS), Cambridge, MA, USA, 7–12 December 2015; pp. 2017–2025.
18. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake, UT, USA, 18–23 June 2018; pp. 7132–7141.
19. Woo, S.; Park, J.; Lee, J.; Kweon, I. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
20. Le, N.; Nguyen, K.; Nguyen, A.; Le, B. Global-local attention for emotion recognition. *Neural Comput. Appl.* **2022**, *34*, 21625–21639. [\[CrossRef\]](#)
21. Cao, W.; Feng, Z.; Zhang, D.; Huang, Y. Facial expression recognition via a CBAM embedded network. *Procedia Comput. Sci.* **2020**, *174*, 463–477. [\[CrossRef\]](#)
22. Duta, I.; Liu, L.; Zhu, F.; Shao, L. Improved residual networks for image and video recognition. In Proceedings of the 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 9415–9422.
23. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-NET: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11531–11539.
24. Safont, G.; Salazar, A.; Vergara, L. Multiclass Alpha Integration of Scores from Multiple Classifiers. *Neural Comput.* **2019**, *31*, 806–825. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Salazar, A.; Safont, G.; Vergara, L.; Vidal, E. Graph regularization methods in soft detector fusion. *IEEE Access* **2023**, *11*, 144747–144759. [\[CrossRef\]](#)
26. Guo, Y.; Zhang, L.; Hu, Y.; He, X.; Gao, J. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; pp. 87–102.
27. Wang, K.; Peng, X.; Yang, J.; Lu, S.; Qiao, Y. Suppressing uncertainties for large-scale facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 6897–6906.
28. Vo, T.; Lee, G.; Yang, H.; Kim, S. Pyramid with super resolution for in the wild facial expression recognition. *IEEE Access* **2020**, *8*, 131988–132001. [\[CrossRef\]](#)
29. 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. [\[CrossRef\]](#) [\[PubMed\]](#)
30. She, J.; Hu, Y.; Shi, H.; Wang, J.; Shen, Q.; Mei, T. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 6248–6257.
31. 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 (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 7660–7669.
32. Huang, Q.; Huang, C.; Wang, X.; Jiang, F. Facial expression recognition with grid-wise attention and visual transformer. *Inf. Sci.* **2021**, *580*, 35–54. [\[CrossRef\]](#)
33. Zeng, D.; Lin, Z.; Yan, X.; Liu, Y.; Wang, F.; Tang, B. Face2exp: Combating data biases for facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 20291–20300.

34. Li, Y.; Lu, G.; Li, J.; Zhang, Z.; Zhang, D. Facial expression recognition in the wild using multi-level features and attention mechanisms. *IEEE Trans. Affect. Comput.* **2020**, *14*, 451–462. [[CrossRef](#)]
35. Liao, J.; Lin, Y.; Ma, T.; He, S.; Liu, X.; He, G. Facial expression recognition methods in the wild based on fusion feature of attention mechanism and LBP. *Sensors* **2023**, *23*, 4204. [[CrossRef](#)] [[PubMed](#)]

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.