applied sciences

MDPI

Article

Context Transformer and Adaptive Method with Visual Transformer for Robust Facial Expression Recognition

Lingxin Xiong [1], Jicun Zhang [2,*], Xiaojia Zheng [1] and Yuxin Wang [1]

[1] School of Computer Science and Technology, Dalian University of Technology, Dalian 116000, China; xionglingxin24@163.com (L.X.); zhengxiaojia91@163.com (X.Z.); wyx@dlut.edu.cn (Y.W.)

[2] Neusoft Reach Automotive Technology (Dalian) Co., Ltd., Dalian 116000, China

* Correspondence: zhangjicun89@163.com

Abstract: In real-world scenarios, the facial expression recognition task faces several challenges, including lighting variations, image noise, face occlusion, and other factors, which limit the performance of existing models in dealing with complex situations. To cope with these problems, we introduce the CoT module between the CNN and ViT frameworks, which improves the ability to perceive subtle differences by learning the correlations between local area features at a fine-grained level, helping to maintain the consistency between the local area features and the global expression, and making the model more adaptable to complex lighting conditions. Meanwhile, we adopt an adaptive learning method to effectively eliminate the interference of noise and occlusion by dynamically adjusting the parameters of the Transformer Encoder's self-attention weight matrix. Experiments demonstrate the accuracy of our CoT_AdaViT model in the Oulu-CASIA dataset as (NIR: 87.94%, VL: strong: 89.47%, weak: 84.76%, dark: 82.28%). As well as, CK+, RAF-DB, and FERPlus datasets achieved 99.20%, 91.07%, and 90.57% recognition results, which achieved excellent performance and verified that the model has strong recognition accuracy and robustness in complex scenes.

Keywords: facial expression recognition; CoT; adaptive method; ViT; complex scenes

check for updates

Citation: Xiong, L.; Zhang, J.; Zheng, X.; Wang, Y. Context Transformer and Adaptive Method with Visual Transformer for Robust Facial Expression Recognition. *Appl. Sci.* **2024**, *14*, 1535. https://doi.org/10.3390/app14041535

Academic Editor: João M. F. Rodrigues

Received: 29 January 2024
Revised: 12 February 2024
Accepted: 12 February 2024
Published: 14 February 2024

1. Introduction

Facial expression recognition is an important research direction in the field of computer vision, which has been receiving extensive attention. Through the more accurate recognition of facial expressions, computers can be endowed with deeper emotional understanding, bringing unprecedented possibilities for human-computer interaction. With the continuous development of technology, facial expression recognition has been widely used in many fields such as medical treatment [1,2], teaching [3], lie detection [4], automatic driving [5], and security driving [6].

However, accurately recognizing facial expressions is still a great challenge due to the interclass similarity and intra-class variability of human facial expressions as well as external environmental factors. Interclass similarity refers to the fact that different facial expressions may look very similar between different facial expression due to subtle differences. For example, the facial expressions of anger and disgust are so close in certain features that they are difficult to distinguish. Intra-class variability refers to the fact that the same class of expression data may be composed of facial images from subjects of different ages, genders, races, etc. This variability adds to the complexity of the recognition, e.g., the facial expression of happiness may vary greatly between people, some may have a bright smile, and others may have a slight upturning of the corners of the mouth. In addition to this, face images in a field environment have arbitrary levels of illumination, non-frontal facial angles, and partially occluded areas, all of which will increase the difficulty of facial expression recognition. To mitigate the effects of ambient light changes on facial features, researchers are actively exploring the use of near-infrared devices for expression recognition, researchers have actively explored the use of near-infrared (NIR) devices for expression recognition. However, the development of NIR-based facial expression recognition has been limited due to the high cost of NIR devices, the lack of wearability, and the relatively low quality and resolution of NIR images. These challenges make accurate recognition of facial expressions in complex scenes a difficult task.

With the continuous development of convolutional networks, the facial expression recognition models proposed based on this technique have achieved excellent performance [7–9]. However, the spatial localization of convolutional networks makes it difficult for the model to learn the dependencies between different facial regions, thus limiting the understanding of global facial expressions. To overcome these problems, some improved models based on convolutional networks have also been proposed, such as the attention mechanism [7], graph convolutional networks [10,11], etc., which aim to enhance the model's ability to learn the dependencies between different facial regions, to better capture the global facial expression information and to improve the model's ability to recognize expressions. In recent years, the Transformer architecture [12] has achieved remarkable success in natural language processing tasks, and inspired by its successful application, ViT (Vision Transformer) [13] has been introduced into image classification tasks and achieved remarkable results through the non-local attention mechanism. Researchers have also introduced it into FER tasks, where its global context-awareness capability can help models better understand global facial expression features. However, since ViT lacks spatial information, is more dependent on large-scale data, and requires more computational resources and time for training and inference. In addition, FER datasets are usually small in size, and training FER models using only the Transformer encoder results in models that are difficult to converge and tend to be biased to focus on occluded and noisy regions. Therefore, the direct use of ViT models for expression recognition may be limited.

To mitigate the effects of illumination changes on facial expression features and to better capture contextual correlations between local regions, we take note of the CoT module [14] (Contextual Transformer), which is a module based on self-attentional improvements to learn correlations between different regions by combining both static and dynamic contextual representations as the final output, as a way of making full use of the information in neighboring regions. We believe that by placing the CoT module behind the convolutional network, we can learn more fine-grained local region feature correlations as a way to reduce the interference of lighting changes on feature extraction, improve the robustness of facial expression recognition, and at the same time better serve the inputs of the ViT. Therefore, using the CoT module to connect the CNN and ViT structures can achieve more comprehensive and flexible feature extraction, thus enhancing the model's comprehensive understanding of facial expressions. On the other hand, to remove the interference of occluded regions and noise, we utilize an adaptive learning method in the Transformer encoder to dynamically retain the feature information most relevant to facial recognition, which not only removes extraneous information but also reduces the amount of computation and improves the performance of the model. The above two methods are used to ensure recognition of facial expressions even in complex scenes. Overall, our main contributions are as follows:

1. The CoT module is used to enhance the model's correlation learning between local regions, capture more fine-grained spatial relationships, improve the ability to perceive subtle differences, make the model more adaptable to complex lighting conditions, and improve the robustness of facial expression recognition.
2. In the Transformer encoder, an adaptive learning method is used to dynamically adjust the parameters of the self-attention weight matrix, enabling the model to retain the top K tokens that are most relevant to the classification of expressions, removing noise and occlusion interferences, improving computational efficiency, and saving computational time.
3. Our proposed model achieves satisfactory results on several datasets. Excellent recognition performance is obtained on the Oulu-CASIA dataset (NIR scene and

natural light under three different lighting conditions), CK+, RAF-DB, and FERPlus datasets, which validates the accuracy and robustness of our proposed model.

2. Related Work

2.1. Near-Infrared Facial Expression Recognition

Expression recognition in NIR images or videos is usually little affected by natural lighting, compared to visible light images. It mainly exploits the unique reflective properties of human skin in the near-infrared spectral range rather than relying on color information in visible light. Zhao et al. [15] used LBP-TOP for near-infrared expression recognition with good recognition results under different lighting conditions. Jung et al. [16] used a new integration approach to combine the two models to extract both appearance and geometric features. Wu et al. [17] designed a fusion of local and global features for facial expression recognition in the NIR. Chen et al. [18] used a combination of SENet and convolutional networks to assign weights to different regions for NIR expression recognition. Zhang et al. [19] constructed a convolutional network to compute the similarity of seven expressions to predict 6 NIR emotional expressions. Salim et al. [20] used a transfer learning approach combined with visible light information for facial expression recognition on NIR images using more additional information.

2.2. Visible Light Facial Expression Recognition

Facial expression recognition under natural light is more mature and widely developed than in near-infrared environments. Facial expression recognition under natural light has been widely researched and developed because it is more in line with the needs of practical applications. Traditional expression recognition methods are usually based on image processing or machine learning techniques, in which relevant features in the image, such as texture, shape, etc., are manually selected or automatically extracted to describe the facial expression features, and then the extracted features are classified using traditional classifiers, such as support vector machines or decision trees. Such an approach often performs poorly on field datasets with light variations and diverse facial poses.

With the development of deep learning, models based on convolutional networks are gradually becoming the mainstream framework for facial expression recognition. Convolutional networks automatically learn features such as edges and textures in an image by capturing information from local regions through convolutional operations and gradually extract higher-level features by stacking multiple convolutional layers, as well as using pooling to maintain more important features while reducing dimensionality and computational complexity, which has led to significant results in different computer vision tasks. A series of novel networks [21–24] have been proposed for general-purpose image classification tasks, which have also laid the foundation for facial expression recognition tasks. Considering the spatial localization of convolutional networks, researchers have proposed some methods to enhance the facial expression understanding of convolutional networks, such as the introduction of pyramid structure [25] to reduce the loss of effective information due to the deeper model or multi-scale convolutional mechanisms [26] to improve the ability to capture subtle changes in expressions. Some other researchers have used graph convolutional networks [10,11] to capture information about the expression features of dynamic sequences or to incorporate different attentional mechanisms depending on the task requirements, etc. Minaee et al. [7] proposed an attention-based convolutional network focusing on different parts of the facial image to perform the FER task. Huang et al. [27] used a lattice attention mechanism in the initial learning phase to enhance feature learning for remote deviations between different facial regions. Zhang et al. [28] improve the performance of recognition by proposing consistency loss to prevent the model from remembering noisy labels, making the model more able to focus on the valid parts. UIlah et al. [29] combine three classical classification networks for feature fusion to cope with occlusion and pose change problems. Gómez-Sirvent et al. [30] performed image segmentation in the middle of the network to cope with the problem of low-resolution facial expression recognition in the wild. Xiao et al. [31] adapted to a complex scene by constraining a joint multitasking network to assign global and local information weights. Naveen et al. [32] coped with the problem of occlusion by using Lanczos interpolation to maintain image quality and combined Hopfield and DBN networks for facial expression recognition tasks. In addition to this, several researchers have used neural network search [33] for micro-expression recognition.

With ViT achieving very competitive results in image classification tasks and transformer-related models achieving more efficient performance on various computer vision tasks [34], some researchers have also introduced related model structures to the expression recognition task. Aouayeb et al. [35] utilized the ViT model by adding an SE module at the end of the model to alleviate the problem of lack of training data for ViT, and achieved effective results on expression datasets with a small amount of data. Li et al. [36] proposed MViT, which makes use of two transformer modules, filtering out the background information and occlusion blocks before classifying them. Ma et al. [37] fused feature information extracted by a convolutional network with an attention-selective fusion module and fed it to the Transformer encoder for classification. Xue et al. [38] combined ViT and MAD methods to learn the rich relationships between local regions, thus selectively removing some irrelevant information and enhancing the recognition ability of the model. Immediately after that, a pooling operation is proposed for the convolutional network and ViT together [39] to remove irrelevant information and further reduce the computation of the model. Yao et al. [40] are inspired by Transformer to focus on more effective feature information, by using the fine-tuning method and channel attention module. Jin et al. [41] based on the Swin Transformer and CBAM module to focus on the important features of face images and designed a loss function to automatically discard unrecognizable samples as a way to perform better facial expression recognition.

3. Proposed Method

3.1. Architecture Overview

The overall model CoT_AdapativeViT framework diagram is shown in Figure 1. The framework mainly consists of an IRNet feature extraction network, a CoT module, and an improved Visual Transformer. Among them, IRNet is used to extract local features, The CoT is used as a connection module to learn the correlation of local features in different regions, to improve the ability to perceive subtle differences so that the model can adapt to the recognition of complex scenes, and ViT is used to learn the global features, and in the self-attention mechanism inside the transformer encoder, adaptive learning method is used, to use the similarity scores with class_token as the basis. In the self-attention mechanism inside the transformer encoder, the adaptive learning method is used to dynamically determine the parameter K based on the similarity score with the class_token, so that the parameters of the self-attention weight matrix are retained in top-K, thus pooling the parameters continuously and retaining only the most effective feature information to achieve the removal of the influence of noise and occlusion interference.

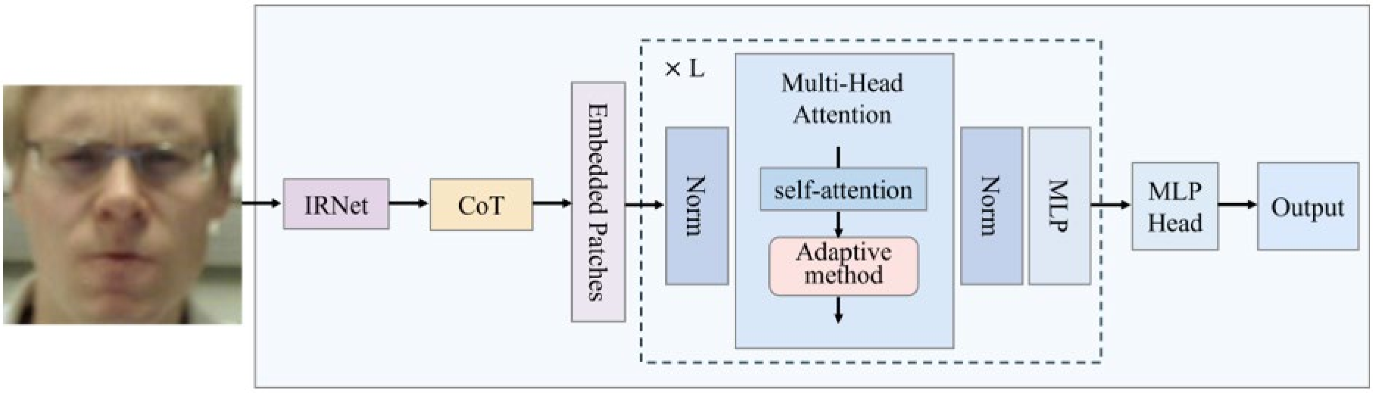


Figure 1. The CoT_AdapativeViT model structure.

3.2. CNN-Based Network

This section describes the CNN network we use in the CoT_AdapativeViT model for local feature extraction. Common image classification networks such as ResNet and EfficientNet are widely used in image classification tasks. ResNet utilizes residual connections to train the deep network, which improves the training of the model. EfficientNet, on the other hand, achieves efficient model structuring by scaling the depth, width, and resolution of the network. These image classification networks have good generalization in many tasks. For the expression recognition task, we further consider the sensitivity of the selected network to face information features. Therefore, we choose the IRNet-50 network and use only the first three basic_blocks as our CNN extraction network. It is shown that IRNet exhibits high recognition performance on different face datasets. This indicates that IRNet has obvious advantages in learning face gesture features and identity-aware features. Using IRNet as our CNN extraction network is expected to improve the performance and robustness of the face expression recognition task based on the enhanced perception of face information.

3.3. Integration Module of CoT between CNN and ViT

In the CoT_AdapativeViT model framework, the CoT module acts as a transition module connecting the CNN feature extraction network and the Transformer encoder. The CoT module can help the model to learn fine-grained correlation information between local regions, to more accurately capture subtle differences in facial expressions and reduce the impact of lighting changes on expression features. By enhancing the correlation of local regions, the perception of expression features can be improved to better cope with the problem of blurred or missing expression features caused by lighting changes.

The CoT module is proposed to make full use of the rich context between neighboring keys as a way to enhance the visual representation ability. In the paper, the CoT module is used to replace the 3 × 3 convolutional network in the ResNet architecture, which demonstrates strong advantages in computer vision tasks such as image recognition and target detection. Unlike the use in the original paper, we use it to connect the CNN and ViT. Through the connection, the local features extracted from the CNN can be better integrated into the overall model, providing richer local information for the subsequent global feature extraction, helping to maintain the consistency between the local area features and the global expression, avoiding the distortion of local features caused by lighting changes, thus improving the stability and accuracy, ensuring the adaptability and robustness of our model in complex scenes.

The overall implementation steps be divided into the overlay work of static context and dynamic context, as shown in Figure 2. Specifically, the input feature map $X \in \mathbb{R}^{H \times W \times C}$, the size of the convolution kernel is defined as K, which is used as the range of the local region. The steps of static context information acquisition are: firstly, use the $K \times K$ convolution kernel to extract the context information, to obtain the static context representation of feature X, which is denoted as $K^1 = Conv_x(X)$; the steps of dynamic context information acquisition are: splice the K^1 obtained from the previous step and X in the channel dimension, to obtain a richer and more diversified representation of the features, and then to enhance the feature representation ability. Then, two consecutive 1×1 convolution operations (one with activation function, denoted as W_θ, and one without activation function, denoted as W_δ), are used to obtain the attention matrix A: $\left[K^1, Q\right] W_\theta W_\delta$, and the weights are normalized using the Softmax function, secondly, the 1×1 convolution is used to operate on the input feature maps X, which is finally multiplied by the normalized weight matrix, thus obtaining K^1 and X. Finally, multiply it by the normalized weight matrix to obtain $K^2 = softmax(A) * (X)$. This step utilizes the additional guidance of the static K^1 to enhance learning capability, so it can dynamically learn the correlation of the local region and obtain dynamic contextual information. Finally, these two are summed to obtain the output of the final CoT module, as shown in Equation (1).

$$CoT(X) = Conv_x(X) + softmax(A) * (X) \tag{1}$$

where A denotes the attention matrix, and $Conv_k$ denotes the convolution operation with convolution kernel size K. In this process, the dimensions of the feature maps before and after the input are kept constant. The value of the convolution kernel K we chose is 3.

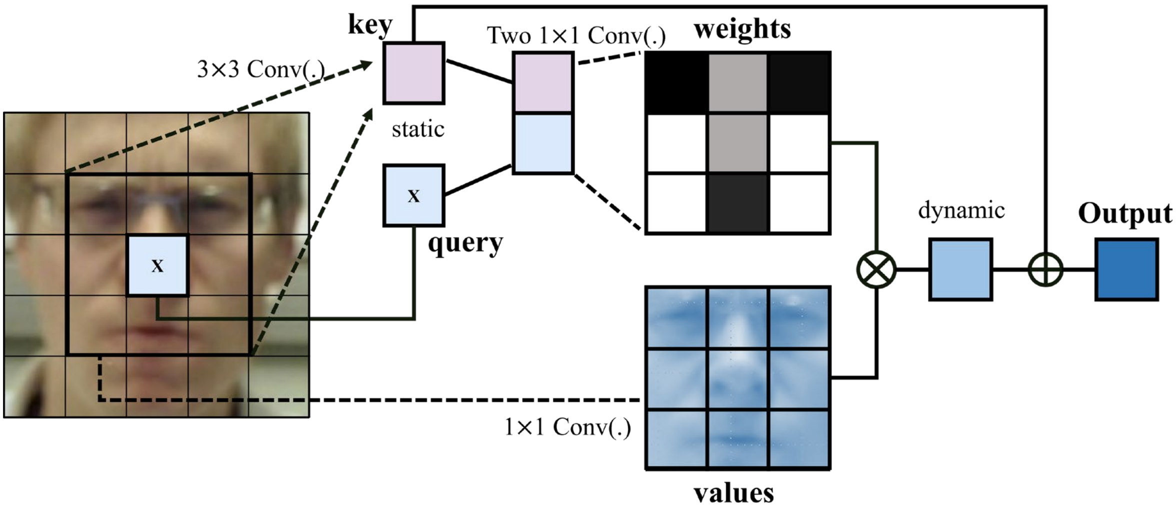


Figure 2. The CoT Module Structure.

By introducing the CoT module, we can make full use of the relationship between local and global features to improve the performance of the facial expression recognition model through the fusion of dynamic and static contexts. This design enables the model to better capture subtle facial features and contextual information and can adapt to the expression recognition task in complex scenes, thus improving the accuracy and robustness of expression recognition.

3.4. Adaptive Method with ViT

Global features are captured using the ViT architecture, and adaptive learning is used in the Transformer encoder architecture to gradually adjust the self-attention weight parameter of each block, which not only focuses on the most relevant features and improves the expressiveness and differentiation of the features, but also suppresses the information that is irrelevant to the task to a certain extent so that it can reduce the attention to noise and redundant information, remove the feature information that is not very relevant to the facial expression, and also effectively reduce the amount of computation and improve the recognition performance of the model.

The whole processing of vit is introduced first. Firstly, the input image is divided into fixed-size image blocks, and, the resulting image vectors are linearly mapped flattened into a vector, and an additional learnable embedding vector class_token is added, thus introducing global semantic information about the whole image and interacting with other image blocks as an important basis for the final classification result. Location information encoding vectors are usually added for these image blocks. The image blocks are concatenated with the class_token and fed into the Transformer Encoder. Each Transformer Encoder consists of Multi-Head Attention, Layer Norm, and MLP modules. These encoders help the model establish global relationships and contextual information in the sequence. Finally, the final recognition result for expression classification is obtained by mapping the class_token feature vector to the category probability distribution by a MLP-Head.

Visual Transformer is built based on the self-attention module, which makes our adaptive token learning approach better able to find more important places to focus on globally, save more effective information, and remove the effects of noise and occlusion. The concept of the self-attention mechanism [9] is to associate different locations of a single sequence to compute a representation of that sequence. Specifically, the input features are divided into three parts: query, key, and value, and for each position of the query, the self-attention mechanism computes the similarity between that query and the keys of all the positions and uses the similarity to assign an attention weight to each position. The attention weights are then weighted and summed with the values of the corresponding location to obtain an output representation of that location. The dot product method is generally used to calculate the similarity, as shown in Equation (2). We use the similarity between other tokens and class_token as a basis to filter out tokens that are not relevant to the global semantic relationship.

$$\left(QK^T\right)_{ij} = \sum_{k=1}^{d} Q_{ik} \cdot K_{jk} \tag{2}$$

After computing the similarity score, it is converted into a set of attention weights that sum to 1 using the Softmax function. These weights are then weighted and summed with the value vector V to get the final attention output as shown in Equation (3), Where Q, K, and V are query vector, key vector, and value vector, respectively. d denotes the dimension of the K vector.

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{3}$$

The main goal of the adaptive learning approach is to determine the number of tokens that need to be retained each time to achieve more effective information filtering in the model's self-attention mechanism. Specifically, we introduce a set of adaptive learning parameters $[K_1, K_2, K_3, \cdots, K_N]$, whose values are dynamically adjusted according to the learning progress of the model during the training process. This parameter is used to specify the number of tokens to be retained during training, and by obtaining the index of the retained elements and using it in conjunction with the attention weight matrix, we can perform the deletion operation on the other elements that are not in K, so that we can focus our attention on the tokens that are the most relevant for the recognition of expressions, and achieve a more efficient pooling and de-noising of the information, thus improving the performance of the model.

4. Experiments

In this section, information specifically related to the experiment will be presented. It includes information about the four datasets used for the experiments (Oulu-CASIA, CK+, RAF-DB, and FERPlus) and the pre-processing procedure of the datasets, followed by a description of the parameter settings during the training process, and finally the ablation experiments and visual analyses and presentations of our proposed model on each dataset.

4.1. FER Datasets

Oulu-CASIA: Contains videos of facial expressions taken by 80 volunteers in a laboratory setting. Each volunteer provided 6 different facial expressions including anger, disgust, fear, happiness, sadness, and surprise. These videos were filmed under different lighting conditions, which can be categorized as dark, weak, or strong according to light intensity. The whole can be divided into shooting in near-infrared and visible light. In this study, the last three frames of each sequence are taken to form the dataset, and experiments are conducted in both cases using 10-fold cross-validation. The dataset images are shown in Figure 3.

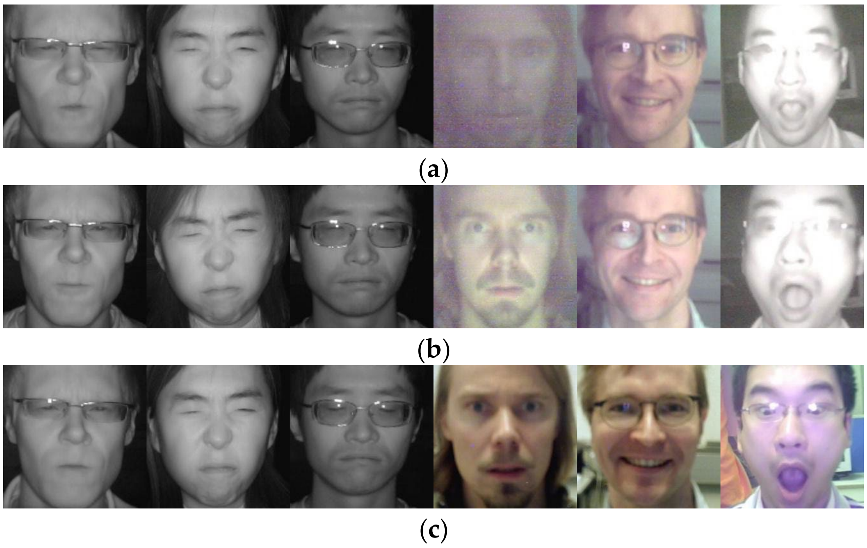

(**a**)
(**b**)
(**c**)

Figure 3. Oulu-CASIA dataset images of different expressions in near-infrared and natural light. (**a**) dark environment (**b**) weak environment (**c**) strong environment.

CK+: Contains 593 facial expression sequences provided by 123 volunteers under laboratory conditions, of which 327 contain accurate expression labels for 6 expressions: anger, disgust, fear, happiness, sadness, and surprise. Each expression sequence varied from neutral to having a distinct expression. In this study, the last three frames of the sequence were selected to form the dataset and experiments were conducted using 10-fold cross-validation methods. The dataset images are shown in Figure 4.

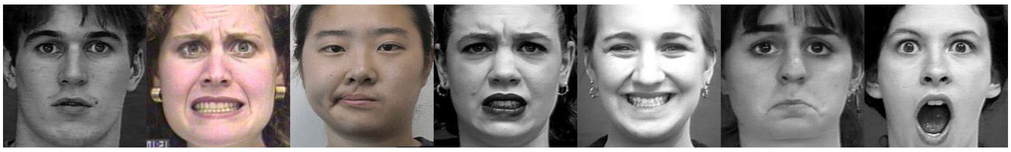


Figure 4. CK+ dataset images of seven different facial expression images.

RAF-DB: Contains 29,672 images of facial expressions of individuals from different languages and cultures, each image is independently labeled by 40 annotators with 7 expressions: anger, disgust, fear, happiness, sadness, surprise, and neutral. In the experiment, 12,271 images were used as training and 3068 images were used for testing. The dataset images are shown in Figure 5.

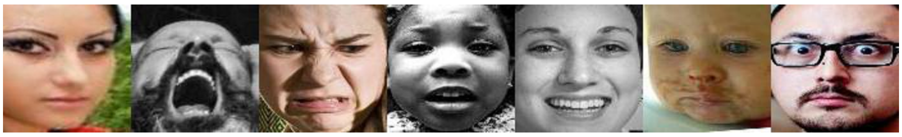


Figure 5. RAF-DB dataset images of seven different facial expression images.

FERPlus: Improved and extended the annotation quality and diversity of the original FER2013 dataset. Labeled by the votes of 10 individuals, removing both unknown and non-facial expression categories, FERPlus contains 8 facial expressions (anger, contempt, disgust, fear, happiness, sadness, surprise, and neutral). In the experiment, a total of 28,557 images were used for training and 3578 for testing. The dataset images are shown in Figure 6.

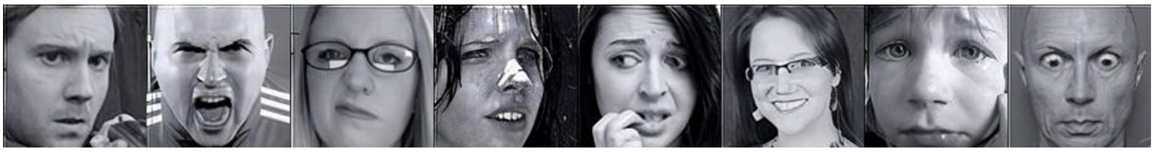


Figure 6. FERPlus dataset images of eight different facial expression images.

4.2. Image Pre-Processing

Most of the face expression datasets include extensive background regions, to improve the computational efficiency and increase the accuracy of recognition of recognition, the face images in the datasets are first detected and cropped, which not only improves the consistency and quality of the training and test data but also ensures that the model focuses on face expression feature extraction and recognition. We perform face detection on four datasets and crop out the face regions as training and testing datasets. The situation is slightly more complicated for the dataset with three different lighting in Oulu-CASIA natural light. The datasets with light intensities of DARK and WEAK have obvious cases of missing details to the extent that the naked eye cannot distinguish whether the images contain humans or not, as shown in Figure 7, and we choose to discard them. The distribution of specific Oulu-CASIA datasets is shown in Table 1.

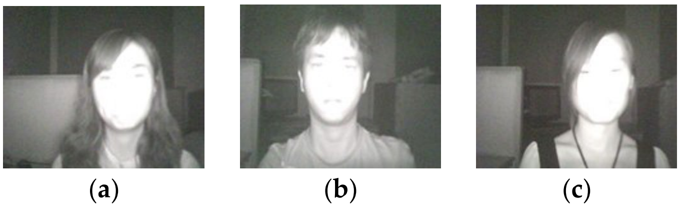


Figure 7. Images that need to be discarded with severe detail loss. (**a**) Discarded image from dark; (**b**) Discarded image from weak; (**c**) Discarded image from dark.

Table 1. The specific number of Oulu-CASIA datasets in visible light.

Luminous Intensity	Pre-Processing Data	Post-Processing Data
strong	1440	1440
weak	1440	1410
dark	1440	1380

4.3. Implementation Details

Models were constructed and implemented using Pytorch. Before inputting into the model, the images were resized uniformly. Considering that the original image size is not consistent across datasets, here, the image of CK+, Oulu-CASIA dataset was resized to 224 × 224 pixels and the image of RAF-DB, FERPlus dataset was resized to 112 × 112 pixels. When reading the dataset, random flip, random rotation, etc. were used for the data enhancement work. The initial value of the adaptive parameter was defaulted to 1. The parameters for training were set as follows: cross-entropy loss function was used, SGD optimizer, the learning rate was fixed to 1×10^{-3}, the batch size was 24 and epoch was set to 100 for CK+, Oulu-CASIA datasets, and the batch size was 128 and epoch was set to 60 for RAF-DB, FERPlus datasets. All experiments were done on a single NVIDI A100 GPU card.

4.4. Ablation Studies

4.4.1. Ablation Experiments on the Proposed Module

To validate the effectiveness of the proposed method, we designed ablation experiments on four datasets, Oulu-CASIA (three different light conditions in near-infrared and visible light), CK+, RAF-DB, and FERPlus. The base model is a hybrid model of IRNet50 loaded with pre-trained weights and a ViT-small framework. Overall, both the CoT module and the adaptive approach can effectively improve the performance of the model. The combination of the two methods can further improve the performance of the model.

Table 2 presents the results of our ablation experiments on the Oulu-CASIA dataset. The results show that our method exhibits excellent recognition performance under both NIR and natural light illumination conditions, showing strong robustness to different scenes. In the NIR dataset, the effects of different lighting variations on the images are relatively small, so we choose to perform the ablation experiments in a dark environment to compare the performance with other models. Compared to the base model, our method improves the accuracy by 2.398%. However, under natural light, the face expression image is greatly affected by lighting variations, which may result in less clear and complete expression features. To verify the adaptability of CoT_AdaptiveViT under different lighting environments, we conducted ablation experiments under the Oulu-CASIA dataset with three lighting conditions. Compared to the base model, we improved the accuracy by 3.31%, 2.66%, and 2.55% under strong, low, and dark lighting conditions, and 1.85% on the NIR dataset, which fully validates the robustness and adaptability of our method in complex scenes.

Table 2. Ablation study on the Oulu-CASIA dataset. The base model refers to the model without the incorporation of the CoT module and the adaptive method.

CoT	**Adaptive Method**	**Visible Light**			**Near-Infrared**
		Strong	**Weak**	**Dark**	**Dark**
×	×	86.16	82.10	79.73	86.09
√	×	87.50	84.25	80.77	86.30
×	√	88.65	82.78	80.02	87.09
√	√	89.47	84.76	82.28	87.94

Table 3 shows the results of our ablation experiments on the CK+, RAF-DB, and FERPlus datasets. Compared to the base model, our method improves by 1.232% on the CK+ dataset, 1.501% in RAF-DB, and 0.806% in FERPlus. The CK+ dataset is a lab-controlled dataset with complete information about the face expressions, and thus, the model achieves a high accuracy of 99.2% on this dataset. The RAF-DB and FERPlus, on the other hand, are field datasets, which face challenges such as light changes, face pose changes, and occlusion. The experimental results show that our model exhibits good solution capabilities in dealing with these problems.

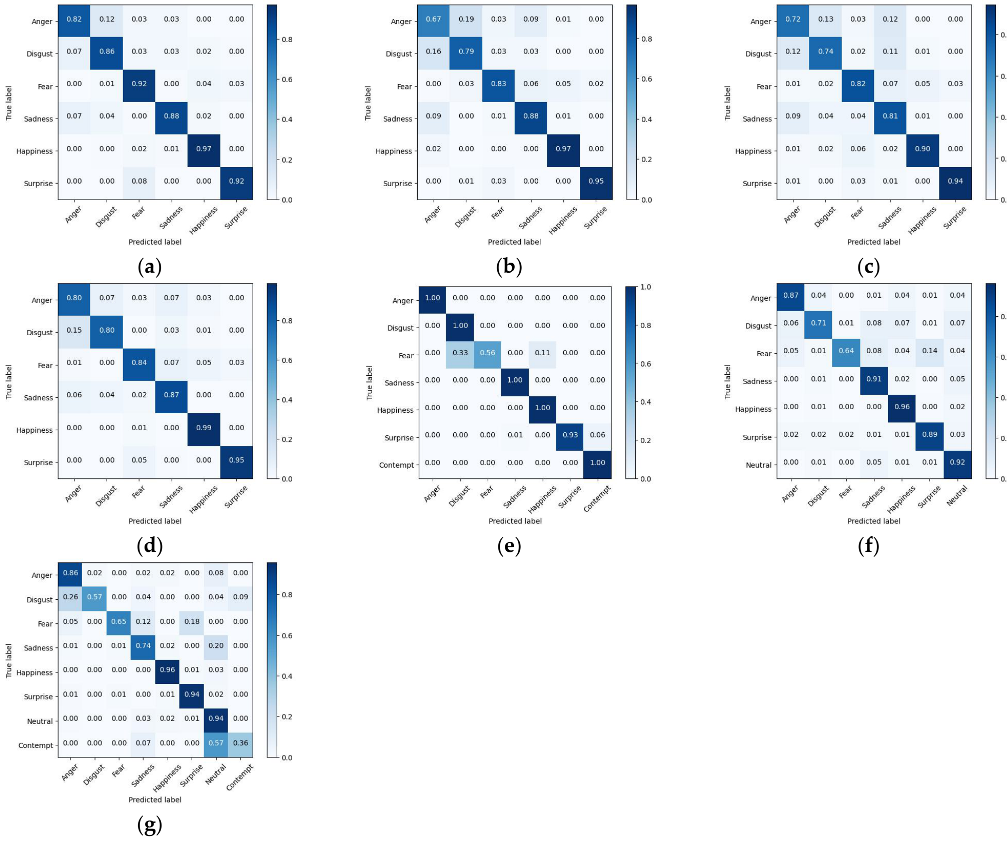
Table 3. Ablation study on the CK+, RAF_DB, and FERPlus datasets. The base model refers to the model without the incorporation of the CoT module and the adaptive method.

CoT	**Adaptive Method**	**CK+**	**RAF-DB**	**FERPlus**
×	×	97.97	89.7	89.76
√	×	98.98	90.61	90.45
×	√	98.71	90.78	90.32
√	√	99.20	91.07	90.57

4.4.2. Visualization and Analysis

Figure 8 shows the results of the visualization of the confusion matrix of the CoT_AdaptiveViT model on the Oulu-CASIA (NIR, VL), CK+, RAF-DB, and FERPlus datasets. From the overall viewpoint, the model is highly discriminative on the categories of Happy, and Surprised and performs well on all other categories. From the recognition results of each category, our proposed model has an excellent global perception ability to adapt to complex scenes. Face expression images under different lighting conditions not only have changes in feature expressions brought about by lighting changes, but are often accompanied by low-resolution images with a lot of noise.

To deeply explore the CoT module's ability to learn the local relevance of face expression features in different environments, we visualized the feature maps before and after the CoT module in the inference stage, as shown in Figure 9. Randomly selected images of face expressions in different environments and with different poses were processed by the CoT module, and we observed that the model focuses more on specific regions of attention, which demonstrates the role of the CoT module in strengthening the model's ability to recognize important features. At the same time, we notice that the local correlations between different regions are enhanced, which helps the model to better capture the subtle changes and emotional information in facial expressions and, mitigate the effects of lighting variations and noise on expression features.



(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

(**g**)

Figure 8. Confusion matrix of CoT_AdaptiveViT on Oulu-CASIA, CK+, RAF-DB, FERPlus. (**a**) Oulu_VL_strong; (**b**) Oulu_VL_weak; (**c**) Oulu_VL_dark; (**d**) Oulu_NIR_dark; (**e**) CK+; (**f**) RAF-DB; (**g**) FERPlus.

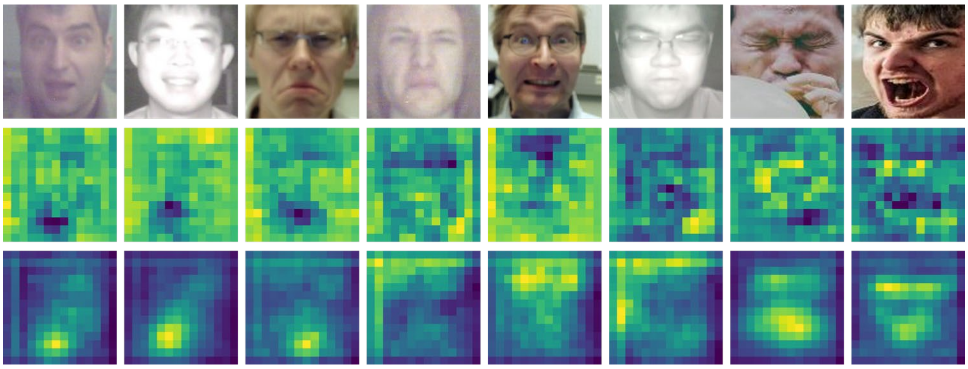


Figure 9. From **top** to **bottom** shows the input image, the feature map before the CoT module, and the feature map after the CoT module.

5. Comparision Study

We use the CoT_AdaptiveViT model to evaluate it in the CK+, RAF-DB, and FERPlus datasets and compare it to state-of-the-art methods.

Results on Oulu-CASIA: The performance comparison of the Oulu-CASIA dataset will be presented in two parts. Firstly, for the full experimental data, we used a 10-fold cross-validation method to verify the recognition performance. For the NIR performance comparison, we achieved state-of-the-art performance as shown in Table 4. Unlike previous single convolutional network models, we used a combination of convolutional network and Transformer Encoder to enhance the comprehensive understanding of facial expressions. Second, in the visible light performance comparison, almost all publicly available models are trained based on datasets that are in a strong light environment. Our model, on the other hand, has been experimented under three different lighting conditions and We will show achieved significant performance gains in all of them, the results of the performance comparison with the existing models under strong lighting environments, as shown in Table 5. Compared to HPFS [42], which uses feature separation as a way to overcome the interference of light variations, our method of local relevance learning using the CoT module is more advantageous.

Table 4. Performance comparison on Oulu-CASIA (NIR dark).

Methods	Year	Acc (%)
LBP-TOP [15]	2011	69.32
DTAGN [16]	2015	66.67
NIRExpNet [17]	2017	78.42
SETFNet [18]	2019	80.34
VGGNet [19]	2020	82.67
Based_transfer learning [20]	2022	84.83
CoT_AdaptiveViT(Ours)	2024	89.74

Table 5. Performance comparison on Oulu-CASIA (VL strong).

Methods	Year	Acc (%)
Spatial-Temporal Network [43]	2017	72.12
HiNet [44]	2019	70.30
VGGNet [19]	2020	84.40
FDRL [45]	2021	88.26
AGFER [46]	2023	89.19
HPFS [42]	2023	87.32
CoT_AdaptiveViT(Ours)	2024	89.47

Results on RAF-DB: Table 6 shows that our accuracy on the RAF-DB dataset is 91.07%, which outperforms other models. EAC [28] uses randomly removing some facial regions to enhance the model's learning of the non-noise part and thus enhances the model's performance. DAN [47] adopts a multi-task joint approach to capture both local and global feature information. Compared to these methods, our proposed method captures local and global features based on the use of adaptive learning to retain tokens that are most relevant to expression categorization, thus achieving the effect of removing noisy and occluded regions and enhancing computational efficiency.

Table 6. Performance comparison on RAF-DB.

Methods	Year	Acc (%)
ViT+SE [35]	2021	87.22
TransFER [38]	2021	90.91
EAC [28]	2022	90.35
RES [48]	2023	90.38
CFNet [31]	2023	87.52
DAN [47]	2023	89.70
CoT_AdaptiveViT(Ours)	2024	91.07

Results on CK+: Table 7 shows that we achieved a high recognition performance on the CK+ dataset by using a 10-fold cross-validation method with 99.20% accuracy. Among them, FER_RN [49] introduces an attention module to redistribute the weight parameters of channel and spatial dimensions, and ZFER [50] performs emotion recognition on faces based on partitions. These methods only take into account local feature information, while our model pairs have a more comprehensive understanding of expression information through the CNN_ViT architecture.

Table 7. Performance comparison on CK+.

Methods	Year	Acc (%)
SL+SSLpuzzling [51]	2021	98.23
FER_RN [49]	2022	96.97
CFNet [31]	2023	99.07
DBN [32]	2023	98.19
CNN_LSTM [52]	2023	92.00
ZFER [50]	2023	98.74
CoT_AdaptiveViT(Ours)	2024	99.20

Results on FERPlus: Table 8 shows that our model achieves 90.57% accuracy on the FERPlus dataset, demonstrating the excellent recognition performance of our proposed model. VTFF [37] uses ViT and feature fusion to capture global-local information, and FT-CSAT [40] fine-tunes the channel-space attentional transformer to allow the model to be more attentive to expression-related information. We use the CoT module to enhance the learning ability of local relations to better understand the correlation between local features, and input the acquired information into the transformer encoder, which enables the model to make full use of local and global information to improve the perception and understanding of expression features.

Table 8. Performance comparison on FERPlus.

Methods	Year	Acc (%)
EAC [28]	2022	89.64
VTFF [37]	2023	88.81
FT-CSAT [40]	2023	89.26
NAGNet [53]	2023	89.30
RRN [54]	2023	89.64
FST [55]	2023	90.41
CoT_AdaptiveViT(Ours)	2024	90.57

6. Conclusions

To address the problem of accurately recognizing facial expressions in complex scenes, we introduce a CoT module between the CNN and VIT architectures, which improves the ability to perceive small differences by learning correlations between features in local regions at a fine-grained level. This enables our model to better adapt to complex lighting variations. Meanwhile, we adopt an adaptive learning method to dynamically adjust the parameters of the self-attentive weight matrix of the converter encoder to better deal with the background occlusion and noise interference problems. Experiments show that our proposed CoT_AdaptiveViT model achieves significant performance improvement on multiple open-source datasets, demonstrating the model's strong recognition accuracy and interference resistance in complex scenes. This also provides new ideas for future research to address the challenges of light changes on expression recognition.

Author Contributions: Data curation, X.Z.; Formal analysis, L.X.; Investigation, X.Z.; Methodology, L.X.; Supervision, J.Z. and Y.W.; Validation, L.X.; Visualization, L.X.; Writing—original draft, L.X.; Writing—review and editing, J.Z. and Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: Author Jicun Zhang was employed by the company Neusoft Reach Automotive Technology (Dalian) Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Li, B.; Blijd-Hoogewys, E.; Greaves-Lord, K.; Stockmann, L. The early development of emotion recognition in autistic children: Decoding basic emotions from facial expressions and emotion-provoking situations. *Underst. Expr. Interact.* **1978**, *37*, 7–210. [CrossRef] [PubMed]
2. Munsif, M.; Ullah, M.; Ahmad, B.; Sajjad, M.; Cheikh, F.A. *Monitoring Neurological Disorder Patients via Deep Learning Based Facial Expressions Analysis*; Springer: Cham, Switzerland, 2022; pp. 412–423.
3. Kabir, M.R.; Dewan, M.A.A.; Lin, F. Lightweight model for emotion detection from facial expression in online learning. In Proceedings of the 2023 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), Regina, SK, Canada, 24–26 September 2023; pp. 174–179.
4. Solbu, A.; Frank, M.G.; Xu, F.; Nwogu, I.; Neurohr, M. The Effect of Synchrony of Happiness on Facial Expression of Negative Emotion When Lying. *J. Nonverbal Behav.* **2023**, *17*, 1–20. [CrossRef]
5. Jain, D.K.; Dutta, A.K.; Verdú, E.; Alsubai, S.; Sait, A.R.W. An automated hyperparameter tuned deep learning model enabled facial emotion recognition for autonomous vehicle drivers. *Image Vis. Comput.* **2023**, *133*, 104659. [CrossRef]
6. Hijji, M.; Yar, H.; Ullah, F.U.M.; Alwakeel, M.M.; Harrabi, R.; Aradah, F.; Cheikh, F.A.; Muhammad, K.; Sajjad, M. FADS: An Intelligent Fatigue and Age Detection System. *Mathematics* **2023**, *11*, 1174. [CrossRef]
7. Minaee, S.; Minaei, M.; Abdolrashidi, A.J.S. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors* **2021**, *21*, 3046. [CrossRef] [PubMed]
8. Shahzad, H.; Bhatti, S.M.; Jaffar, A.; Akram, S.; Alhajlah, M.; Mahmood, A. Hybrid Facial Emotion Recognition Using CNN-Based Features. *Appl. Sci.* **2023**, *13*, 5572. [CrossRef]
9. 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]
10. Zhou, J.; Zhang, X.; Liu, Y. Learning the connectivity: Situational graph convolution network for facial expression recognition. In Proceedings of the 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP), Macau, China, 1–4 December 2020; pp. 230–234.
11. Xu, C.; Du, Y.; Wang, J.; Zheng, W.; Li, T.; Yuan, Z. A joint hierarchical cross-attention graph convolutional network for multi-modal facial expression recognition. *Comput. Intell.* **2023**. [CrossRef]
12. 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*. [CrossRef]
13. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
14. Li, Y.; Yao, T.; Pan, Y.; Mei, T. Contextual transformer networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1489–1500. [CrossRef] [PubMed]
15. Zhao, G.; Huang, X.; Taini, M.; Li, S.Z.; PietikäInen, M.J.I. Facial expression recognition from near-infrared videos. *Image Vis. Comput.* **2011**, *29*, 607–619. [CrossRef]
16. Jung, H.; Lee, S.; Yim, J.; Park, S.; Kim, J. Joint fine-tuning in deep neural networks for facial expression recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2983–2991.

17. Wu, Z.; Chen, T.; Chen, Y.; Zhang, Z.; Liu, G.J.A.S. NIRExpNet: Three-stream 3D convolutional neural network for near infrared facial expression recognition. *Appl. Sci.* **2017**, *7*, 1184. [CrossRef]
18. Chen, Y.; Zhang, Z.H.; Zhong, L.; Chen, T.; Chen, J.X.; Yu, Y.D. Three-Stream Convolutional Neural Network with Squeeze-and-Excitation Block for Near-Infrared Facial Expression Recognition. *Electronics* **2019**, *8*, 385. [CrossRef]
19. Zhang, Z.; Lai, C.; Liu, H.; Li, Y.-F.J.N. Infrared facial expression recognition via Gaussian-based label distribution learning in the dark illumination environment for human emotion detection. *Neurocomputing* **2020**, *409*, 341–350. [CrossRef]
20. Salim, N.R.; Srinath, V.; Jayaraman, U.; Gupta, P. Recognition in the near infrared spectrum for face, gender and facial expressions. *Multimed. Tools Appl.* **2022**, *81*, 4143–4162. [CrossRef]
21. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
22. 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, 1–26 July 2016; pp. 770–778.
23. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 1–26 July 2016; pp. 2818–2826.
24. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; pp. 6105–6114.
25. Fang, J.; Lin, X.; Liu, W.; An, Y.; Sun, H. Triple attention feature enhanced pyramid network for facial expression recognition. *J. Intell. Fuzzy Syst.* **2023**, *44*, 8649–8661. [CrossRef]
26. Lan, J.; Jiang, X.; Lin, G.; Zhou, X.; You, S.; Liao, Z.; Fan, Y. Expression recognition based on multi-regional coordinate attention residuals. *IEEE Access* **2023**, *11*, 63863–63873. [CrossRef]
27. Huang, Q.; Huang, C.; Wang, X.; Jiang, F.J.I.S. Facial expression recognition with grid-wise attention and visual transformer. *Inf. Sci.* **2021**, *580*, 35–54. [CrossRef]
28. Zhang, Y.; Wang, C.; Ling, X.; Deng, W. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 418–434.
29. Ullah, Z.; Mohmand, M.I.; Rehman, S.U.; Zubair, M.; Driss, M.; Boulila, W.; Sheikh, R.; Alwawi, I. Emotion recognition from occluded facial images using deep ensemble model. *Comput. Mater. Contin.* **2022**, *73*, 4465–4487.
30. Gómez-Sirvent, J.L.; López de la Rosa, F.; López, M.T.; Fernández-Caballero, A. Facial Expression Recognition in the Wild for Low-Resolution Images Using Voting Residual Network. *Electronics* **2023**, *12*, 3837. [CrossRef]
31. Xiao, J.; Gan, C.; Zhu, Q.; Zhu, Y.; Liu, G. CFNet: Facial expression recognition via constraint fusion under multi-task joint learning network. *Appl. Soft Comput.* **2023**, *141*, 110312. [CrossRef]
32. Naveen, P. Occlusion-aware facial expression recognition: A deep learning approach. *Multimed. Tools Appl.* **2023**, 1–27. [CrossRef]
33. Verma, M.; Mandal, M.; Reddy, S.K.; Meedimale, Y.R.; Vipparthi, S.K. Efficient neural architecture search for emotion recognition. *Expert Syst. Appl.* **2023**, *224*, 119957. [CrossRef]
34. Bobojanov, S.; Kim, B.M.; Arabboev, M.; Begmatov, S.J.A.S. Comparative Analysis of Vision Transformer Models for Facial Emotion Recognition Using Augmented Balanced Datasets. *Appl. Sci.* **2023**, *13*, 12271. [CrossRef]
35. Aouayeb, M.; Hamidouche, W.; Soladie, C.; Kpalma, K.; Seguier, R. Learning vision transformer with squeeze and excitation for facial expression recognition. *arXiv* **2021**, arXiv:2107.03107.
36. Li, H.; Sui, M.; Zhao, F.; Zha, Z.; Wu, F. MVT: Mask vision transformer for facial expression recognition in the wild. *arXiv* **2021**, arXiv:2106.04520.
37. Ma, F.; Sun, B.; Li, S. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Trans. Affect. Comput.* **2021**, *14*, 1236–1248. [CrossRef]
38. Xue, F.; Wang, Q.; Guo, G. Transfer: Learning relation-aware facial expression representations with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3601–3610.
39. Xue, F.; Wang, Q.; Tan, Z.; Ma, Z.; Guo, G. Vision transformer with attentive pooling for robust facial expression recognition. *IEEE Trans. Affect. Comput.* **2022**, *14*, 3244–3256. [CrossRef]
40. Yao, H.; Yang, X.; Chen, D.; Wang, Z.; Tian, Y. Facial Expression Recognition Based on Fine-Tuned Channel–Spatial Attention Transformer. *Sensors* **2023**, *23*, 6799. [CrossRef]
41. Jin, Z.; Zhang, X.; Wang, J.; Xu, X.; Xiao, J. Fine-Grained Facial Expression Recognition in Multiple Smiles. *Electronics* **2023**, *12*, 1089. [CrossRef]
42. Yang, L.; Yang, H.; Hu, B.B.; Wang, Y.; Lv, C. A Robust Driver Emotion Recognition Method Based on High-Purity Feature Separation. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 15092–15104. [CrossRef]
43. Zhang, K.; Huang, Y.; Du, Y.; Wang, L. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Trans. Image Process.* **2017**, *26*, 4193–4203. [CrossRef] [PubMed]
44. Verma, M.; Vipparthi, S.K.; Singh, G. Hinet: Hybrid inherited feature learning network for facial expression recognition. *IEEE Lett. Comput. Soc.* **2019**, *2*, 36–39. [CrossRef]
45. 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.

46. Arabian, H.; Battistel, A.; Chase, J.G.; Moeller, K. Attention-Guided Network Model for Image-Based Emotion Recognition. *Appl. Sci.* **2023**, *13*, 10179. [CrossRef]
47. Wen, Z.; Lin, W.; Wang, T.; Xu, G. Distract your attention: Multi-head cross attention network for facial expression recognition. *Biomimetics* **2023**, *8*, 199. [CrossRef]
48. Lin, Z.; She, J.; Shen, Q. Real emotion seeker: Recalibrating annotation for facial expression recognition. *Multimed. Syst.* **2023**, *29*, 139–151. [CrossRef]
49. Jiang, Q.; Peng, X.; Chen, H.; Guo, Y. Facial expression recognition based on residual network. In Proceedings of the 2022 41st Chinese Control Conference (CCC), Hefei, China, 25–27 July 2022; pp. 7000–7006.
50. Shahzad, T.; Iqbal, K.; Khan, M.A.; Iqbal, N. Role of zoning in facial expression using deep learning. *IEEE Access* **2023**, *11*, 16493–16508. [CrossRef]
51. Pourmirzaei, M.; Montazer, G.A.; Esmaili, F. Using self-supervised auxiliary tasks to improve fine-grained facial representation. *arXiv* **2021**, arXiv:2105.06421.
52. Mohana, M.; Subashini, P.; Krishnaveni, M. Emotion Recognition from Facial Expression Using Hybrid cnn–lstm Network. *Int. J. Pattern Recognit. Artif. Intell.* **2023**, *37*, 2356008. [CrossRef]
53. Zhu, H.; Hu, P.; Tang, X.; Xia, D.; Huang, H. NAGNet: A novel framework for real-time students' sentiment analysis in the wisdom classroom. *Concurr. Comput. Pract. Exp.* **2023**, *35*, e7727. [CrossRef]
54. Jiang, C.-S.; Liu, Z.-T.; Wu, M.; She, J.; Cao, W.-H. Efficient facial expression recognition with representation reinforcement network and transfer self-training for human–machine interaction. *IEEE Trans. Ind. Inform.* **2023**, *19*, 9943–9952. [CrossRef]
55. Feng, H.; Huang, W.; Zhang, D.; Zhang, B. Fine-tuning swin transformer and multiple weights optimality-seeking for facial expression recognition. *IEEE Access* **2023**, *11*, 9995–10003. [CrossRef]

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.