Information Sciences 580 (2021) 35–54

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Facial expression recognition with grid-wise attention and visual transformer

Qionghao Huang [a,b], Changqin Huang [a,b,*], Xizhe Wang [a], Fan Jiang [a]

[a] Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Jinhua, 321004 Zhejiang, China
[b] School of Information Technology in Education, South China Normal University, Guangzhou, 510631 Guangdong, China

ARTICLE INFO

Article history:
Received 9 May 2021
Received in revised form 9 July 2021
Accepted 12 August 2021
Available online 21 August 2021

Keywords:
Facial Expression Recognition
Attention mechanism
Grid-wise attention
Visual transformer

ABSTRACT

Facial Expression Recognition (FER) has achieved remarkable progress as a result of using Convolutional Neural Networks (CNN). Relying on the spatial locality, convolutional filters in CNN, however, fail to learn long-range inductive biases between different facial regions in most neural layers. As such, the performance of a CNN-based model for FER is still limited. To address this problem, this paper introduces a novel FER framework with two attention mechanisms for CNN-based models, and these two attention mechanisms are used for the low-level feature learning the high-level semantic representation, respectively. In particular, in the low-level feature learning, a grid-wise attention mechanism is proposed to capture the dependencies of different regions from a facial expression image such that the parameter update of convolutional filters in low-level feature learning is regularized. In the high-level semantic representation, a visual transformer attention mechanism uses a sequence of visual semantic tokens (generated from pyramid features of high convolutional layer blocks) to learn the global representation. Extensive experiments have been conducted on three public facial expression datasets, CK+, FER+, and RAF-DB. The results show that our FER-VT has achieved state-of-the-art performance on these datasets, especially with a 100% accuracy on CK + datasets without any extra training data.

1. Introduction

Facial Expression Recognition (FER), which classifies the emotion on images [21], has become a hot research topic in the field of computer vision in recent years. Convolutional Neural Networks (CNNs) based FER models also have achieved a remarkable performance as other domains do, such as object detection, image segmentation, and image generation. However, each convolutional filter in a CNN operates only on a small region. This spatial locality makes a model difficult to learn the structural dependencies between different facial units in most neural layers. Therefore, CNN-based FER models capture only facial features, without understanding facial expression images globally.

There are two categories of solutions to mitigate this problem in the literature. The first category is to increase kernel sizes [1], increase model depth [17], or adopt new operations like global pooling [21]. The approaches in this category attempt to overcome the weaknesses of convolution locality by increasing computational complexity without addressing the long-range learning problem well. Another category is about integrating CNNs with long-range inductive biases. Long-range inductive biases are usually learned by using attention mechanisms (e.g Transformer [40]) or sequential models

[*] Corresponding author at: Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Jinhua, 321004 Zhejiang, China.

https://doi.org/10.1016/j.ins.2021.08.043
0020-0255/© 2021 Elsevier Inc. All rights reserved.

(e.g. LSTM [42]). These models encode the temporal or spatial dependency relationships (e.g. time-series data, natural language sentences) between different data points, such as RAN [41], and DACL [11]. To some extent, these hybrid models have overcome the shortcoming of convolution filters, achieving impressive performance over pure CNN-based models. However, they neglect the fact that convolution filters in different layers perform different levels of feature extraction [30].

As shown in Fig. 1, convolutional filters in the stem layer play a critical role in learning lower local features such as textures, and edges [30]. These features are crucial for performing a fined-grain recognition task, such as FER [7]. FER models must learn a global understanding of the facial image at this very first stage by paying more attention to the important local features. Research efforts are made to solve this problem, like M-CRT [9], DPND [6], and the recent work on facial motion prior networks (FMPN) [7]. In particular, FMPN utilizes a facial motion mask network to generate an attention facial mask with long-range inductive biases learned from facial muscle motions. The attention map produces a fused image to regularize the parameter learning in extracting the low-level features. But the mechanism employed in the facial motion mask network is vulnerable to pose and occlusion variant settings. Therefore, a more robust method for attention map generation is urgent to promote the performance of a FER.

Fig. 1 shows that the filters in deeper level convolutional layers focus on learning a high abstract semantic feature representation [34,11]. Research works in the literature have shown that different feature fusion strategies (e.g. BiFPN [36]) or attention mechanisms (e.g. ASFF [22]) on these features contribute to the performance of CNN-based models. The features obtained from these deeper convolutional filters can be regarded as abstract semantic representations, just like the tokens of natural languages [43]. As such, a transformer-based mechanism [40] may be appropriate for compensating the convolutional filters of deeper level layers in learning long-range inductive biases. With a strong capacity in learning long-range inductive biases for computer vision tasks, visual transformer [13] techniques like ViT [8] need to learn spatial inductive biases from a large-scale dataset (over 300 million images for ViT) to learn spatial inductive biases to achieve state-of-the-art performance. However, such huge datasets are usually not available for a FER task. Therefore, without relying on a large-scale FER dataset, how to retain the powerful modeling capacity of visual transformer in learning long-range inductive biases to overcome the weakness of CNN-based models for FER remains a challenging issue to be addressed.

In summary, most existing FER methods focus only on using the attention mechanism either for the low-level feature extraction [35] or for the deeper features [11]. In addition, just one identical attention mechanism is used for the whole model [27]. As such, these attention mechanisms cannot achieve the best performance for CNN-based FER models.

To overcome the aforementioned shortcomings, in this paper, we propose a novel framework with two attention mechanisms (noted as FER-VT) for CNN-based FER models at both the low-level feature learning stage and high semantic representation stage. In the low-level feature learning stage, Grid-Wise Attention (GWA) is designed to model the long-range dependencies among different regions of a facial expression image. In particular, there are three main blocks in this part:

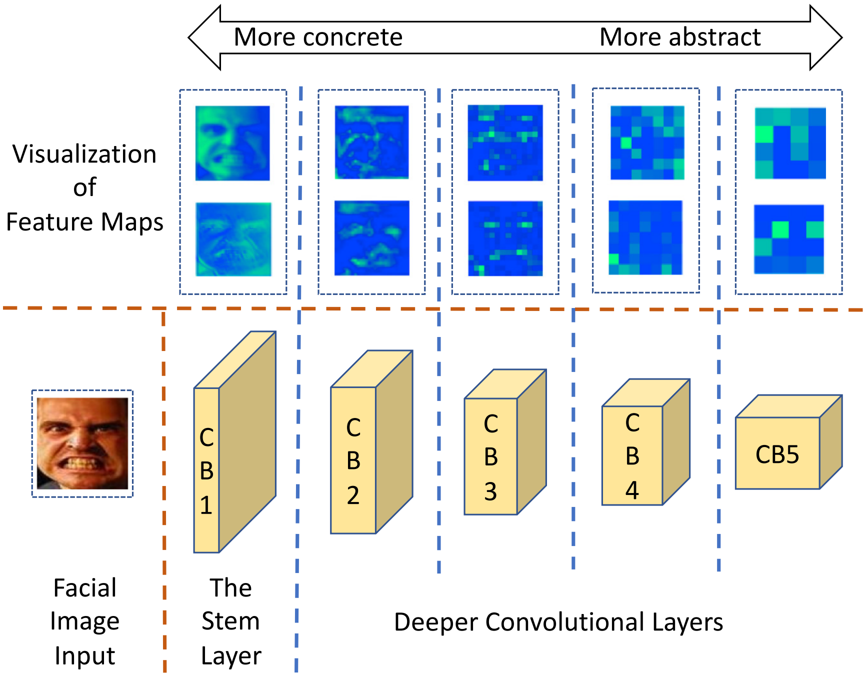


Fig. 1. Visualization of feature maps from different convolutional blocks of a convolutional network – a Resnet, which consists of five convolutional neural blocks. The stem layer performs a low-level feature extraction task like handcrafted features (e.g. Gabor [5], LBP [32]). The deeper convolutional layers focus on learning more abstract feature learning, like word tokens of natural languages. Note that CB stands for Convolutional neural Block.

local feature extraction, grid-wise attention calculation, and residual feature fusion. The learnable parameters in the convolutional filters are regularized to learn better features for the FER task with these blocks. At the high semantic representation stage, a Visual Transformer Attention (VTA) is applied to a sequence of visual semantic tokens. These visual tokens are generated from the pyramid features of a candidate backbone network. The global representation vectors with spatial inductive biases from convolutional filters and long-range dependency inductive biases from VTA are generated by this component and forwarded to a facial expression classifier. To the best of our knowledge, it is the first time to employ a token-based visual transformer technique in FER tasks. Extensive experiments have been conducted on three public facial expression recognition benchmark datasets, CK+, FERplus, and RAF-DB. The results show that our method has superior performance compared to existing methods, especially with **a 100% accuracy** on CK+ datasets **without any extra training data**.

Our main contributions are summarized as follows:

- We propose a novel FER framework (FER-VT) with two attention mechanisms that overcome the weakness of CNN-based models in learning long-range inductive biases for enhancing the performance of performing a FER task.
- We design a grid-wise attention mechanism that uses long-range dependencies between different facial regions to regularize the convolutional parameter learning in the low-level feature extraction for a FER task.
- We also introduce a token-based visual transformer to learn long-range inductive biases in high-level semantic feature learning. To the best of our knowledge, this is the first to apply a token-based visual transformer technique for the FER task.
- We conduct the compared experiments on three public FER benchmark datasets. The results demonstrate that FER-VT has achieved state-of-the-art performances. And we also do an ablation study that specifically demonstrates the effectiveness of each component in our model.

The remainder of this paper is organized as follows: Section 2 reviews the related work. Section 3 explains the proposed method of FER-VT in detail. Section 4 reports the experiment results, followed by the conclusion in Section 5.

2. Related work

In this section, we review relevant research works on facial expression recognition (FER) from the following three perspectives.

Traditional Methods for FER. Traditional researches usually utilize handcrafted features or shallow learning to perform a FER task [38]. Bazzo et al. [5] introduced a Gabor feature-based facial recognition system where the Gabor kernels were used on facial expression subtracted from an averaged neutral face. Some work [32] also indicates that LBP features are effective and efficient for facial expression recognition. Geometry-based features are also effective for FER tasks, such as, a facial action units (AUs) framework that directly models a facial activation for specific facial expressions[37]. Some hybrid features (e.g. texture and geometry) were also proposed [39]. Zhong et al. [48] developed an SVM-based sparse learning method for expression analysis. However, since 2013, some facial recognition competitions such as FER2013 [12] and EmotiW [21] have collected facial expressions relatively close to real-world scenarios, which advance the research transition of FER from lab-controlled settings to in-the-wild ones. These works based on handcrafted features or shallow learning are vulnerable to an in-the-wild setting, resulting in an unsatisfactory performance on these challenging datasets [21].

CNN-based Models FER. As convolutional neural networks (CNN) can extract deeper and more spatial inductive biases information, CNN-based models have gained their prevalence in facial expression recognition tasks [19,15,21,47]. Bargal et al. [3] proposed a hybrid model which combines VGG16 with a residual neural network to learn facial surface features of expressions. Kumar et al. [19] used CNN-based Kinect APIs to extracted 71 facial points to represent facial expressions for gesture recognition. Shao et al. [33] combined three kinds of convolutional neural networks, i.e., shallow CNN, dual branch CNN, and transfer-learning-based CNN, for robust facial expression recognition in the wild. Hossain et al. [15] used a pretrained CNN model and two deep sparse auto-encoders to extract facial and speech features, and employed a support vector machine to determine a corresponding emotion label under a secure edge and cloud computing environment. Mohan et al. [28] designed a local gravitational force descriptor for local features from face images, and the descriptor is fed into a deep convolution neural network model to perform the FER task, with achieving a sound performance. Considering both emotion label information and local spatial distribution information of samples simultaneously, Zheng et al. [47] proposed a discriminative multi-task learning method for facial expression recognition. Also, some works are explored toward video-based methods with sound performance. Zhao et al. [46] also proposed a weighted mechanism in a sequence of image frames to define the peak and non-peak frames for the final classification. To learn dynamic-still information in a sequence of facial frames, Zhang et al. [44] introduced a spatial–temporal CNN-based network. These CNN-based models achieved a sound performance compared to traditional methods, but the weakness of convolutional filters in learning long-range dependency biases has limited their sounder performance.

Attention Mechanism in FER. Attention mechanisms have been widely used in natural language processing (NLP) [40]. Due to their efficiency in modeling the dependencies between different features, they are also applied to CV tasks in recent years, e.g. visual transformer [18]. Recent works also explore their applications in the FER task [41,11]. Minaee et al. [27] proposed an attention-based convolutional network, focusing on different parts of a facial image to perform the FER task.

Liu et al. [23] proposed an attention mechanism in hierarchical scales to discover the most relevant regions to the facial expression and select the most informative scales to learn the expression-discriminative representations. Wang et al. [41] designed a region attention network (RAN) to capture the importance of facial regions for occlusion and pose variant FER. Farzaneh et al. [11] proposed a deep attentive center loss method. Their model employs an attention mechanism to estimate attention weights correlated with feature importance by using the intermediate spatial feature maps in CNN as a context. These attention-based hybrid models have mitigated the weakness of convolutional filters in learning long-range inductive biases and enlarged the receptive field [18]. However, as previously discussed, most of these models did not take the properties of convolutional filters in different feature learning stages into considerations. A more fined-grain attention mechanism is essential to promote the performance of the CNN-based model for FER.

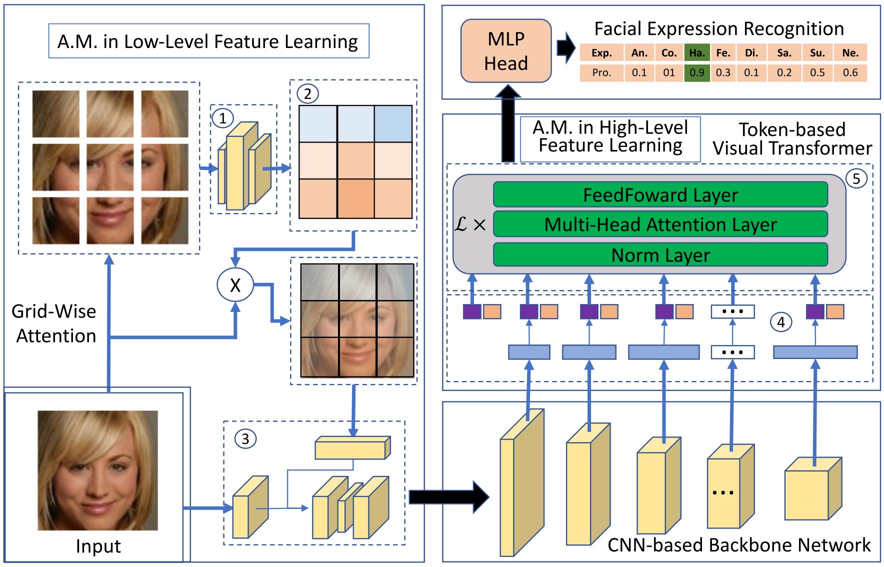
3. The proposed method

3.1. Framework of FER-VT

The framework is depicted in Fig. 2. The framework mainly consists of three parts: Backbone, attention mechanism in low-level feature learning, and attention mechanism in high-level feature learning. The figure also shows an implementation of the framework (noted as FER-VT) where Grid-Wise Attention (GWA) acts as the attention mechanism in low-level feature learning and Visual Transformer Attention(VTA) for the attention mechanism in high-level feature learning. All components employed in the model are differentiable, thus, FER-VT is a fully end-to-end model for facial expression recognition. The three main components of FER-VT are briefly described in the following.

- **Backbone Network**. This component can employ recent popular deep learning models, such GoogleNet, ResNet, or EfficientNet as a backbone network.
- **Grid-Wise Attention**. This component consists of **Local Feature Extraction**, **Grid-Wise Attention Calculation**, and **Residual Feature Fusion**. With these sub-components, the long-range biases between different facial regions can be learnt by using convolutional filters in the stage of low feature learning (detailed in SubSection 3.3).
- **Visual Transformer Attention**. This component consists of **Visual Token Generation**, and **Token-based Visual Transformer**. The high semantic features learned from high-level convolutional filters are transformed into a sequence of visual tokens. VTA will learn a global representation for FER tasks (detailed in SubSection 3.4).

In the following, we first describe each part of **Backbone Network**, **Grid-Wise Attention** and **Visual Transformer Attention** in details, then elaborate on the learning process of the whole model.


A.M. in Low-Level Feature Learning
①
②
X
Grid-Wise Attention
Input
③
MLP Head
Facial Expression Recognition
A.M. in High-Level Feature Learning
Token-based Visual Transformer
⑤
FeedFoward Layer
Multi-Head Attention Layer
Norm Layer
$\mathcal{L}\times$
④
CNN-based Backbone Network

Exp.	An.	Co.	Ha.	Fe.	Di.	Sa.	Su.	Ne.
Pro.	0.1	01	0.9	0.3	0.1	0.2	0.5	0.6

Fig. 2. The framework of FER-VT. Notations: ① Local Feature Extraction, ② Grid-Wise Attention Calculation, ③ Residual Feature Fusion, ④ Visual Token Generation, and ⑤ Visual Transformer. Abbreviation: A.M. for Attention Mechanism, Exp. for Expressions, Pro. for Probability, An. for Anger, Co. for Contempt, Ha. for Happiness, Fe. for Fear, Di. for Disgust, Sa. for Sadness, Su. for Surprise, and Ne. for Neutral.

3.2. Backbone network

In this section, we introduce an implementation of the backbone network. As Resnet [14] is very popular in image classification tasks, we also use Resnet as the backbone for FER-VT. Table 1 shows the network structure of the backbone network, which is derived from Kaiming et al. [14], and the average pool, flatten, and a fully connected network are removed after the *conv5_x block*.

We denote the collection of the feature maps after each convolutional layer as pyramid features as other researchers do. The grid-wise attention is designed for the low-level feature learning of facial expressions, and the mechanism will employ the long-range inductive biases to regularize the parameter learning of the stem layer (detailed in Section 3.3). The feature maps in the pyramid feature collection will be resized to fixed-length visual semantic tokens, to which a visual transformer attention is applied to learn a global representation for the facial expression (detailed in Section 3.4).

3.3. Attention mechanism in low-level convolutional filters

In this section, we introduce an attention mechanism for convolutional filters in low-level feature learnings, i.e., the grid-wise attention, which is used to mitigate the weakness (in learning long-range biases in the low-level feature extraction of facial expression images) of convolutional filters. There are mainly three parts in the GWA: local feature extraction, grid-wise attention calculation, and residual feature fusion. The details are presented in the following.

3.3.1. Local feature extraction

Formally, an aligned facial image is denoted as $I^{C\times H\times W}$, where C is the channel of the image (3 for an RGB mode), H is the height, and W is the width. The image will be cropped into $h \times w$ grids before being forwarded to the feature extraction network. The collection of these grids is denoted as:

$$Grid(I,h,w) = \left\{ I_{1,1}^{C\times\frac{H}{h}\times\frac{W}{w}}, \ldots, I_{i,j}^{C\times\frac{H}{h}\times\frac{W}{w}}, \ldots, I_{h,w}^{C\times\frac{H}{h}\times\frac{W}{w}} \right\}, = I^{hw\times C\times\frac{H}{h}\times\frac{W}{w}}, \tag{1}$$

where h and w represents the input image I is divided into $h \times w$ grids, $I_{i,j}^{C\times\frac{H}{h}\times\frac{W}{w}}$ ($I_{i,j}$ for short) represents this image grid is with a shape of $C \times \frac{H}{h} \times \frac{W}{w}$, and locates in the ith row and the jth column in the image grids of I ($1 \leqslant i \leqslant h, 1 \leqslant j \leqslant w$, and $i,j \in \mathbb{N}$).

Each grid $I_{i,j}$ will be forwarded to a local feature extraction network to learn the local feature of facial regions in the grid. We employ an inverted bottleneck neural block to build the local feature extraction network. The structure of the local feature extraction network is shown in Table 2, the number of the channel of $I_{i,j}$ varies in the network as the sequence $c \rightarrow ck \rightarrow c$, like an inverted bottleneck. As we want to learn a low feature from different facial regions for an attention calculation, the shape of these grids of a facial image remains unchanged after the local feature network, and comparisions are made between different settings to show the effectiveness of this design in Section 4.6. We denote the low feature extraction process as:

$$\widehat{I}_{i,j} = LFN(I_{i,j}), \widehat{I}^{hw\times C\times\frac{H}{h}\times\frac{W}{w}} = LFN\left(I^{hw\times C\times\frac{H}{h}\times\frac{W}{w}}\right), \tag{2}$$

where $\widehat{I}_{i,j}$ is the feature map obtained from the $I_{i,j}$ using the low feature extraction network. These feature maps are forwarded to the grid-wise attention calculation to weight their importance to a FER task.

Table 1
Structure of the backbone network for FER-VT (The structure is derived from Kaiming et al. [14]).

Layer name	feature map size	Resnet18	Resnet34	Resnet50	Resnet101	Resnet152
conv1 (stem layer)	$\mathfrak{h} \times \mathfrak{w}$	7×7, 64, stride 2				
conv2_x	$\frac{\mathfrak{h}}{2} \times \frac{\mathfrak{w}}{2}$	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, & 64 \\ 3\times3, & 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, & 64 \\ 3\times3, & 64 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, & 64 \\ 3\times3, & 64 \\ 1\times1, & 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, & 64 \\ 3\times3, & 64 \\ 1\times1, & 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, & 64 \\ 3\times3, & 64 \\ 1\times1, & 256 \end{bmatrix} \times 3$
conv3_x	$\frac{\mathfrak{h}}{4} \times \frac{\mathfrak{w}}{4}$	$\begin{bmatrix} 3\times3, & 128 \\ 3\times3, & 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, & 128 \\ 3\times3, & 128 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, & 128 \\ 3\times3, & 128 \\ 1\times1, & 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, & 128 \\ 3\times3, & 128 \\ 1\times1, & 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, & 128 \\ 3\times3, & 128 \\ 1\times1, & 512 \end{bmatrix} \times 8$
conv4_x	$\frac{\mathfrak{h}}{8} \times \frac{\mathfrak{w}}{8}$	$\begin{bmatrix} 3\times3, & 256 \\ 3\times3, & 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, & 256 \\ 3\times3, & 256 \end{bmatrix} \times 6$	$\begin{bmatrix} 1\times1, & 256 \\ 3\times3, & 256 \\ 1\times1, & 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1\times1, & 256 \\ 3\times3, & 256 \\ 1\times1, & 1024 \end{bmatrix} \times 23$	$\begin{bmatrix} 1\times1, & 256 \\ 3\times3, & 256 \\ 1\times1, & 1024 \end{bmatrix} \times 36$
conv5_x	$\frac{\mathfrak{h}}{16} \times \frac{\mathfrak{w}}{16}$	$\begin{bmatrix} 3\times3, & 512 \\ 3\times3, & 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, & 512 \\ 3\times3, & 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, & 512 \\ 3\times3, & 512 \\ 1\times1, & 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, & 512 \\ 3\times3, & 512 \\ 1\times1, & 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, & 512 \\ 3\times3, & 512 \\ 1\times1, & 2048 \end{bmatrix} \times 3$

Table 2
Structure of the local feature extraction network.

Input	Operator	Output
$C \times \frac{H}{h} \times \frac{W}{w}$	Conv2d, Kernel 1 × 1, Stride 1	$(Ck) \times \frac{H}{h} \times \frac{W}{w}$[†]
$Ck \times \frac{H}{h} \times \frac{W}{w}$	BatchNorm	$(Ck) \times \frac{H}{h} \times \frac{W}{w}$
$Ck \times \frac{H}{h} \times \frac{W}{w}$	LeakyReLU	$(Ck) \times \frac{H}{h} \times \frac{W}{w}$
$Ck \times \frac{H}{h} \times \frac{W}{w}$	Conv2d, Kernel 1 × 1, Stride 1	$C \times \frac{H}{h} \times \frac{W}{w}$
$C \times \frac{H}{h} \times \frac{W}{w}$	BatchNorm	$C \times \frac{H}{h} \times \frac{W}{w}$
$C \times \frac{H}{h} \times \frac{W}{w}$	LeakyReLU	$C \times \frac{H}{h} \times \frac{W}{w}$

[†] k is the expansion factor.

3.3.2. Grid-Wise Attention Calculation

The grid-wise attention takes the features from different facial regions into considerations, addressing the previously discussed weakness of convolutional filters in the low-level feature learning stage.

The first step in this block is to calculate the similarity between different grids using a matrix dot production operator. The process is formulated as:

$$d_k = \frac{W}{w}, query = \widehat{I}^{hw \times C \times \frac{H}{h} \times \frac{W}{w}}, key = \widehat{I}^{hw \times C \times \frac{W}{w} \times \frac{H}{h}}, scores = \frac{query \times key}{d_k}, attn = softmax(scores), = \widehat{I}^{hw \times C \times \frac{H}{h} \times \frac{H}{h}}, \tag{3}$$

where × between matrices represents a matrix dot product.

The attention mechanism employed in this part is slightly different from the self-attention mechanism. We use adaptive pooling to squeeze each channel into a scalar after an attention mechanism, and expand the channel back to the original shape. The process is formulated as:

$$aap = AdaptiveAvgPool2d((1,1)), \widehat{I}^{hw \times C \times 1 \times 1} = aap(attn), expander = Ones\left(\frac{H}{h}, \frac{W}{w}\right), pattn = expander * \widehat{I}^{hw \times C \times 1 \times 1},$$
$$= \widetilde{I}^{hw \times C \times \frac{H}{h} \times \frac{W}{w}} \tag{4}$$

where $*$ represents the scalar matrix product between matrices with a broadcasting property, *AdaptiveAvgPool2d* with the parameter $(1,1)$ denotes an adaptive pooling technique that converts an operand matrix into a scalar, and *Ones* with the parameter $(\frac{H}{h}, \frac{W}{w})$ is to generate a matrix with all elements being equal to 1 in the shape of $\frac{H}{h} \times \frac{W}{w}$.

With the obtained attention weight for each grid *pattn*, a global attention matrix is formed by concatnating these weights back to the shape of the original matrix.

$$\widetilde{I}^{C \times H \times W} = Ungrid\left(\widetilde{I}^{hw \times C \times \frac{H}{h} \times \frac{W}{w}}\right), \widetilde{I'}^{C \times H \times W} = \widetilde{I}^{C \times H \times W} * I^{C \times H \times W} \tag{5}$$

where *Ungrid* is an inverse operation in Eq. (1), which puts these grid attentions back to the shape of the original facial image *I*.

Thus, $\widetilde{I'}^{C \times H \times W}$ ($\widetilde{I'}$ for short) is a feature map with the consideration of long-range biases between different facial regions in the low-level feature learning stage.

3.3.3. Residual feature fusion

We apply a residual network technique to fuse the features between the original image I and the weighted feature map $\widetilde{I'}$ for the backbone network.

The feature fusion network has a similar shape as the letter ‘Y’, i.e., three blocks: two feature transformation networks and one feature fusion network. These two feature transformation networks share the structure, and the specification is presented in Table 3.

The original image I and the weighted feature map $\widetilde{I'}$ are forwarded to their feature transformation network, respectively, where the learning parameters in these feature transformation networks are not shared. After the feature transformation, they are forwarded to the fusion network. The process is formulated as follows:

Table 3
Structure of the feature transformation networks.

Input	Operator	Output
$C \times H \times W$	Conv2d, Kernel 3 × 3, Padding 1, Stride 1	$C \times H \times W$
$C \times H \times W$	BatchNorm	$C \times H \times W$
$C \times H \times W$	ReLU	$C \times H \times W$

$$\bar{I}^{C\times H\times W} = RFN\left(FT1(I) + FT2\left(\widetilde{I}\right)\right) \tag{6}$$

where *FT1* denotes the feature transformation network for the original facial image, *FT2* denotes the feature transformation network for the weighted feature map $\widetilde{I}'$, and *RFN* denotes the residual feature fusion network. The specification of the structure of the fusion network aree 4.

So far, the low-level feature extraction of facial images with the grid-wise attention mechanism has been described in this section. The obtained feature map $\bar{I}^{C\times H\times W}$ is forwarded to the candidate backbone network.

3.4. Attention mechanism in high-level convolutional filters

The attention mechanism described in the previous section focuses on feature extraction in the low-level convolutional filters. In this section, we introduce an implementation of the attention mechanism for high-Level convolutional filters, i.e. visual transformer attention(VTA), which is used to enhance the long-range biases learning in convolutional filters at the high level. There are mainly two parts in the VTA: visual token generation, and token-based visual transformer. The details are presented in the following.

3.4.1. Visual token generation

As a visual transformer takes tokens as input like recurrent networks [42], we first generate visual semantic tokens from the backbone network. With two steps of pyramid feature extraction and embedding of tokens. As the convolutional filters have learned the spatial biases, the requirement of the size of the training dataset can be reduced significantly. We will demonstrate the advantage in the experiment section.

Pyramid Feature Extraction. Similar to pyramid feature networks, we extract the pyramid features from the high-level convolutional blocks of the candidate backbone network. Because the shapes of feature maps are not identical, the first must be resize the feature maps to the same shape. We then employ convolutional neural blocks to resize the feature maps from the backbone network. The specification of the pyramid feature transformation networks is reported in Table 5. Since the shape of the input feature maps is not identical, we transform them into the same shape as $C_1 \times H_1 \times W_1$, denoting their subscript notation as $*$. Suppose that three feature maps are extracted from the backbone network, its process can be formulated as: Given three feature maps are extracted the backbone network, the process can be formulated as:

$$L_1'^{C_1\times H_1\times W_1} = PFT1\left(L_1''^{C_*\times H_*\times W_*}\right), L_2'^{C_1\times H_1\times W_1} = PFT2\left(L_2''^{C_*\times H_*\times W_*}\right), L_3'^{C_1\times H_1\times W_1} = PFT3\left(L_3''^{C_*\times H_*\times W_*}\right), \tag{7}$$

where *PFT1*, *PFT2*, *PFT3* denote the feature transformation network from the pyramid features L_1'', L_2'', and L_3'', respectively. Their learnable parameters are not shared.

Visual Token Embedding. The resized feature maps (e.g. L_1', L_2' and L_3') are transformed to visual semantic tokens so as to fit the input requirement to a visual transformer. These feature maps will reshape into fixed-length vectors, that is (take 3 feature maps for example):

$$L_i^{1\times C_1 H_1 W_1} = reshape\left(L_i'^{C_1\times H_1\times W_1}\right),\ i = 1, 2, 3 \tag{8}$$

where L_i is a visual token for the feature map L_i'. We employ a fully connected network with a high generalization to perform the embedding of visual tokens. The embedding of visual tokens can be formulated as follow:

$$T_i^{1\times D} = token\_embed\left(L_i^{1\times C_1 H_1 W_1}\right),\ i = 1, 2, 3 \tag{9}$$

where *token_embed* is the visual token embedding network. Its structure is shown in Table 6, and the learnable parameters are shared to all token embeddings.

Table 4
Structure of the residual feature fusion network.

Input	Operator	Output
$\widetilde{I}^{C\times H\times W}$	Conv2d, Kernel 3 × 3, Padding 1, Stride 1, Use-bias	$C \times H \times W$
$C \times H \times W$	BatchNorm, PReLU	$C \times H \times W$
$C \times H \times W$	Conv2d, Kernel 3 × 3, Padding 1, Stride 1, Use-bias	$C \times H \times W$
$C \times H \times W$	BatchNorm, PReLU	$\widetilde{I}'^{C\times H\times W}$
$\widetilde{I}^{C\times H\times W} + \widetilde{I}'^{C\times H\times W}$	Conv2d, Kernel 3 × 3, Padding 1, Stride 1	$C \times H \times W$
$C \times H \times W$	Sigmoid	$\bar{I}^{C\times H\times W}$

Table 5
Structure of the feature resize networks.

Input	Operator	Output
$C_* \times H_* \times W_*$	Conv2d, Kernel $k \times k$, Padding $\frac{k-1}{2}$, Stride s	$C_1 \times H_1 \times W_1$
$C_1 \times H_1 \times W_1$	BatchNorm	$C_1 \times H_1 \times W_1$
$C_1 \times H_1 \times W_1$	LeakyReLU	$C_1 \times H_1 \times W_1$

3.4.2. Token-based Visual Transformer

As previously discussed, a pixel-level visual transformer requires a large-scale dataset to learn local spatial biases as convolutional filters do. Without requiring huge training datasets, our visual semantic token-level visual transformer has the capacity of a CNN-based backbone in learning long-range biases in the high-level semantic feature learning of a facial image.

As the standard Transformer needs a *CLS* token to the input sequence, we use a similar technique in ViT [8], prepend a learnable embedding *CLS* token T_{cls} the input sequence. Thus, the visual token sequence of a facial image is:

$$(T_{cls}, T_1, \ldots, T_i, \ldots, T_N), \ i = 1, 2, \ldots, N \tag{10}$$

where T_i is a visual token with the dimension of D.

A standard learnable 1-dimensional position embedding is added to the visual token embeddings to retain positional information about pyramid features. The resulting sequence of visual tokens serves as input to the encoder of a standard Transformer. The visual token sequence with positional embeddings is denoted as:

$$Z_0 = (T_{cls}; T_1, \ldots, T_i, \ldots, T_N) + E_{pos}, \ i = 1, 2, \ldots, N \tag{11}$$

We apply the same encoder as ViT [8], which mainly consists of repeated blocks of (multi-head self-attention mechanism) MHSA and MLP. Layernorm, residual connections and GELU non-linearity techniques are applied in these blocks. The process is formulated as follow:

$$z'_{\mathfrak{l}} = MHSA(z_{\mathfrak{l}-1}) + z_{\mathfrak{l}-1}, \mathfrak{l} = 1, 2, .., \mathfrak{L}, z_{\mathfrak{l}} = MLP(z'_{\mathfrak{l}}) + z'_{\mathfrak{l}}, \mathfrak{l} = 1, 2, .., \mathfrak{L}, \tag{12}$$

where $\mathfrak{L}$ is the number of the repeated blocks of MHSA and MLP, noting that every fully-connected network in each block is applied with a layer-normed technique to avoid an overfitting problem.

We use $z^0_{\mathfrak{L}}$ as the global vector for the FER task with a layer-normed fully-connected network, and the process is denoted as follows:

$$y = FER(LayerNorm(z^0_{\mathfrak{L}})) \tag{13}$$

where the linear neural block *FER* is in the shape of $D \times K$, and K is the number of facial expressions that need to be classified.

3.5. Model process and model fitting

With descriptions in the previous subsections, we are ready to present the workflow of FER-VT with details on the training loss.

Model Process. Fig. 2 shows the whole process of FER-VT. Some notations are defined as follows to describe the model process. The grid-wise attention neural block is denoted as $GWA(\cdot)$), the candidate backbone as $Backbone(\cdot)$), the position embeddings as $POS(\cdot)$, the visual transformer attention as $VTA(\cdot)$), the loss calculation as $\mathscr{L}(\cdot)$, and the prediction as $FER(\cdot)$. By using these notations, the feedforward process of the model is summarized in Algorithm 1.

Algorithm 1. Facial Expression Recognition with GWA and VTA

```
Input: {$I^{B\times C\times H\times W}$, $h$, $w$, $K$} # $B$ is the batchsize, $h$ and $w$ is the grid parameter, $K$ is the number of classes.
Output: {$Y^B$} # the predicted facial expression labels.
$\bar{I}^{B\times C\times H\times W}$ = $GWA(I^{B\times C\times H\times W}, h, w)$ # refer to SubSection 3.3
$L'^{B\times N\times C_1\times H_1\times W_1}$= $Backbone(\bar{I}^{B\times C\times H\times W})$ # refer to Eq. (7)
$L^{B\times N\times C_1H_1W_1}$= $reshape(L'^{B\times N\times C_1\times H_1\times W_1})$ # refer to Eq. (8)
$T^{B\times N\times D}$= $token\_embed(L^{B\times N\times C_1H_1W_1})$ # refer to Eq. (9)
$Z_0^{B\times(N+1)\times D}$= $POS(T^{B\times N\times D})$ # refer to Eq. (11)
$Z_{\mathfrak{L}}^{B\times D}$= $VTA(Z_0^{B\times(N+1)\times D})$ # refer to Eq. (12)
$Y^{B\times K}$ = $FER(Z_{\mathfrak{L}}^{B\times D})$ # refer to Eq. (13)
return {$Y^{B\times K}$}
```

Table 6
Structure of the visual token embedding network.

Input	Operator	Output
$1 \times C_1H_1W_1$	LayerNorm	$1 \times C_1H_1W_1$
$1 \times C_1H_1W_1$	Relu	$1 \times C_1H_1W_1$
$1 \times C_1H_1W_1$	Droupt	$1 \times C_1H_1W_1$
$1 \times C_1H_1W_1$	Linear	$1 \times D$

Model Fitting. The learnable parameters in FER-VT come from four parts: the *GWA* block, the *Backbone* block, the *VTA* block, and the *FER* block. We use a label smoothing in the loss function to reduce overfitting. Thus, the loss is expressed as:

$$loss = -\frac{\sum_{i=1}^{B} log\left(softmax\left(y_i^{pred}\right)\right) * smoothed\_labels\left(y_i^{gth}\right)}{B} \tag{14}$$

where B is the batchsize in a feedforward process of the model, *log* is a logarithmic function, and *smoothed_labels* is a label smoothing function. In particular, the function of *smoothed_labels* is defined as:

$$smoothed\_labels(y_{hot}) = y_{hot}(1 - \alpha) + \frac{\alpha}{K} \tag{15}$$

where y_{hot} is a one-hot encoding, and α is a hyperparameter (usually equal to 0.1).

As FER-VT is fully differentiable, we can use a stochastic gradient descent algorithm (denoted as $SGD(\cdot)$) to minimize the objective function.

4. Experiments

In this section, we report our experiments on three widely used datasets, CK+ [24], FER+ [45], and RAF-DB [29]. We compare our model with state-of-the-art methods and do an ablation study to demonstrate the effectiveness of each component in our model.

4.1. Datasets

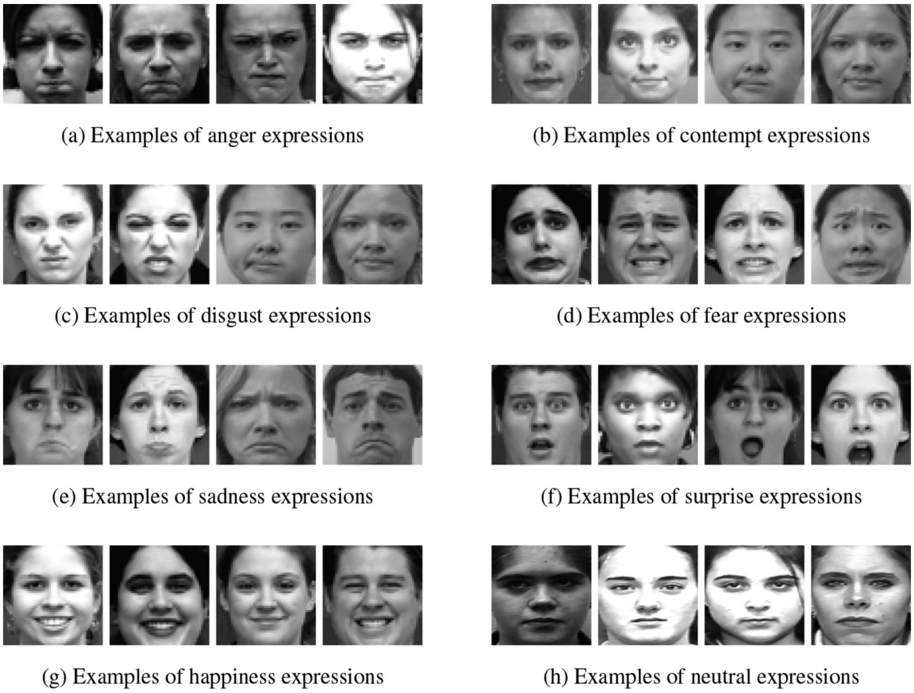
We employ the most popular 10-fold cross-validation in our experiments on the following three datasets.

CK+ Dataset. The Extended Cohn-Kanade Dataset [24] is a laboratory-controlled benchmark dataset labeled with seven basic expressions (anger, contempt, disgust, fear, happiness, sadness, and surprise) and a neutral expression. Some example expressions in the dataset are shown in Fig. 3. Following the settings of [7] to extract the last three frames of each labeled sequence, we obtain the resulting dataset with 948 images of seven basic expressions and 948 images of a neutral expression. We denote the resulting dataset with six basic expressions (except contempt ones and neutral ones) as CK+6 (900 images), seven basic expressions (except neutral ones) as CK+7 (948 images), and the one with eight expressions as CK+8 (1896 images).

FER+Dataset. The FER+dataset [4] is a facial expression dataset in the wild, and its original version is FER2013 dataset [12]. But the labeling accuracy of FER2013 is not reliable. Some example expressions in the dataset are shown in Fig. 4. Ten people were assigned to manually label the basic emotion for each image in the dataset. The final emotion label was assigned based on the votes from the ten people, the resulting dataset with eight facial expressions (anger, contempt, disgust, fear, happiness, sadness, surprise, and neutral) with 22082 images in total (17656 images training, 2189 for validation, 2237 for testing). The distributions of facial expressions for training, testing, and validation on the FER+ are roughly the same.

RAF-DB Dataset. Shang et al. [20] provide the Real-world Affective Faces Database (RAF-DB) for facial expression recognition, which contains approximately 30,000 facial expression images with uncontrolled poses and illuminations from thousands of individuals of diverse ages and races. Each image is independently labeled by approximately 40 annotators. The dataset consists of two subsets: the single-label subset (7-class basic emotions) and the two-tab subset (11-class compound expressions). The single-label subset with 7-class basic emotions is used for the evaluation in this paper. In particular, these are anger, disgust, fear, happiness, sadness, surprise, and neutral. Some example expressions in the dataset are shown in Fig. 5. From this subset, we use 12,271 face-aligned images for the training set and 3,068 for the test set.

Occlusion and Pose Variant Datasets. To verify whether FER-VT is robust to occlusion and variant pose issues on in-the-wild settings, we employ the datasets that are collected by the work [41] from the FER+ dataset and the RAF-DB dataset, in which the facial expression images are of occlusion or variant pose. These datasets are noted as Occlusion-RAF-DB, Pose-RAF-DB, Occlusion-FER+, and Pose-FER+, respectively.


(a) Examples of anger expressions

(b) Examples of contempt expressions

(c) Examples of disgust expressions

(d) Examples of fear expressions

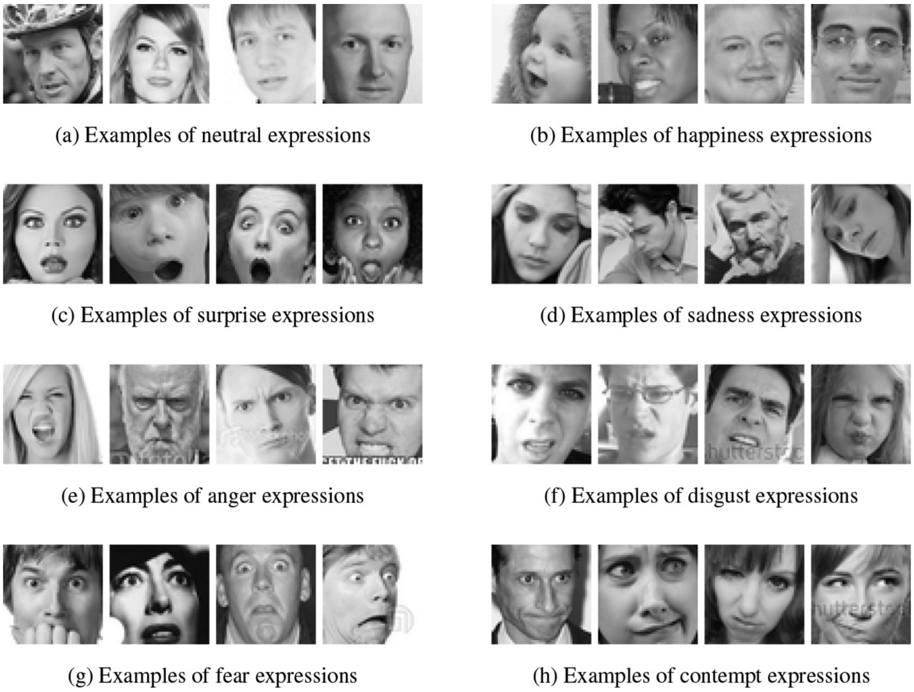
(e) Examples of sadness expressions

(f) Examples of surprise expressions

(g) Examples of happiness expressions

(h) Examples of neutral expressions

Fig. 3. Some expression examples of CK+ dataset.


(a) Examples of neutral expressions

(b) Examples of happiness expressions

(c) Examples of surprise expressions

(d) Examples of sadness expressions

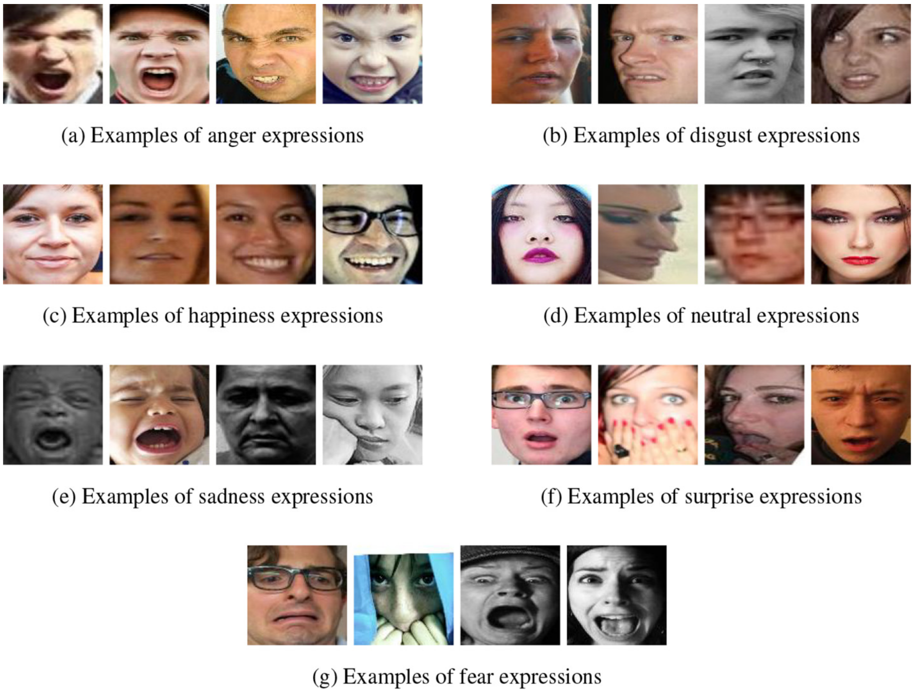
(e) Examples of anger expressions

(f) Examples of disgust expressions

(g) Examples of fear expressions

(h) Examples of contempt expressions

Fig. 4. Some expression examples of FER+ dataset.



(a) Examples of anger expressions

(b) Examples of disgust expressions

(c) Examples of happiness expressions

(d) Examples of neutral expressions

(e) Examples of sadness expressions

(f) Examples of surprise expressions

(g) Examples of fear expressions

Fig. 5. Some expression examples of RAF dataset.

4.2. Experiment setup

In this section, we present the detailed specification of FER-VT and the setup for its training.

4.2.1. Model specification

The channels in the inverted bottleneck neural block in *GWA* change as 3 → 64 → 3, while they stay the same as 3 in the residual feature fusion block during training, this is following the setting of the work [31]. A Resnet [14] is used as the backbone network of FER-VT, and the feature maps from the last three blocks (i.e. $N = 3$) of the backbone are used to generate visual semantic tokens ($D = 128$). Thus, the length of the input sequence to VTA is 4. As for the VTA, we set the depth as 12, the heads of multi-head attention as 8, and the dim of the project key vector as 64, and these are following the setting of ViT [8]. The enhancement techniques such as BatchNorm, LayerNorm, Label Smoothing are also applied in the model.

4.2.2. Training setup

The initialization of weights and biases in *Residual Feature Fusion* in Section 3.3.3 uses *Xavier* uniform distribution with *gain*=0.1. *Kaiming* uniform distribution is used to initialize weights and biases of the rest convolutional filters in the model. The initialization of weights and biases in the attention blocks follows the corresponding settings in ViT. The negative scope of the activation LeakyReLU in all neural blocks is set as 0.1. The above settings are common initialization strategies in the domains. Furthermore, we train the model using the Adam optimization with an adaptive learning rate strategy, and its initial learning rate is 0.001. The mini-batch size is 32, the dropout rate is 0.2, and the learning factor is set to 0.1. Also, these are common settings for the Adam optimizer. Some common data augmentation techniques are also employed in the experiments.

4.3. Comparisons to state-of-the-art methods

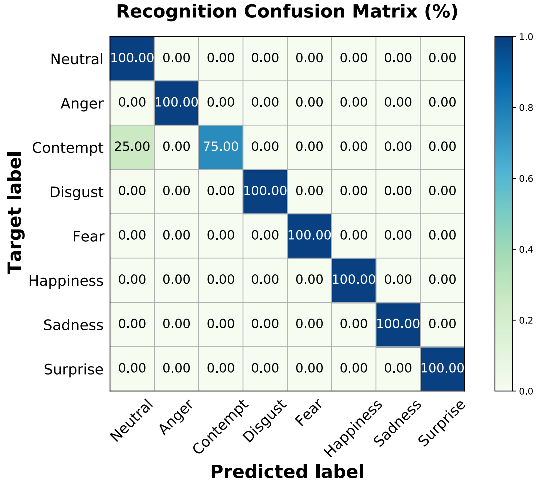
We use CK+ [24], FER+ [45], and RAF-DB [29] datasets for evaluation. We compare our method with state-of-the-art approaches in terms of prediction accuracy.

Results on CK+. As the results shown in Table 7, our method FER-VT for the first time achieves a **100% accuracy** on CK+ dataset of 6 or 7 facial expressions, and **99.46% accuracy** for 8 facial expressions. The results outperforms the compared state-of-the-art both in video-based and image-based methods. Although FAN and FER-GCN are video-based FER models, which extract the frames of images to boost the performance of classification, FER-VT still outperforms them. As shown in Fig. 6, FER-VT only wrongly recognized some *contempt* expressions as the *neutral* ones. These contempt expressions may also be recognized as *neutral* ones by humans. Overall, FER-VT also achieves an impressive performance in the 8-

Table 7
Prediction accuracy of the evaluated models on the test dataset of CK+.

Models	Accuracy		
	6 classes	7 classes	8 classes
CNN+DSAE; SVM [15]	98.27%	–	–
DDMTL [47]	97.63%	–	–
FER-GCN [10]	–	99.54%	–
FMPN [7]	–	98.06%	–
ROI-CNN [35]	–	–	94.67%
FAN [26]	–	99.70%	–
ViT [8]†	97.59%	96.88%	96.21%
FER-VT (Ours)†	**100.00%**	**100.00%**	**99.46%**

*The second-best performance.
† Without a pretrained model.



	Neutral	Anger	Contempt	Disgust	Fear	Happiness	Sadness	Surprise
Neutral	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Anger	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
Contempt	25.00	0.00	75.00	0.00	0.00	0.00	0.00	0.00
Disgust	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00
Fear	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
Happiness	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00
Sadness	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00
Surprise	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00

Fig. 6. Confusion matrix for FER-VT in recognizing eight facial expressions on CK+ dataset.

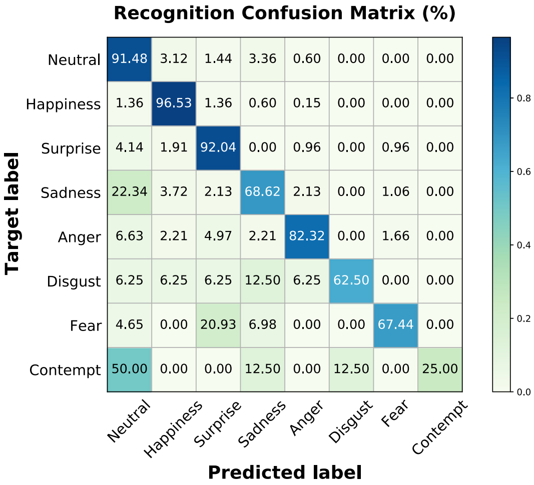
class recognition task even without a pretrained model compared with other state-of-the-art models. One reason for this is that FER-VT considers the long-range inductive biases in both low-level feature learning and high semantic feature representation. Another important reason is that the facial images in the CK+ dataset are collected in a laboratory-controlled environment, which results in high-quality face alignment. This works well with the grid-wise attention in FER-VT. Furthermore, FER-VT did **not need a pretraining on any extra facial dataset** to reach its performance on the CK+ dataset.

Results on FER+. Compared to all the state-of-the-art methods on the FER+ dataset, our model achieves the best performance with a **90.04%** accuracy rate, as shown in Table 8. This is the first time that the accuracy has been achieved above 90% accuracy for this dataset. Specifically, Fig. 7 shows the confusions matrix on the test set of FER-VT: happiness has the highest accuracy of 96.53%, followed by *Surprise* (92.04%), *Neutral* (91.48%), and *Anger* (82.32%). Compared with the other models, the accuracies for *Sadness*, *Disgust* and *Fear* have been improved significantly. The accuracy for *Contempt* is still at a low level (25%), due to an insufficient contempt images for training. As such, FER-VT could not learn to distinguish it from *Neutral*, *Sadness*, or *Disgust*. Some expressions have high chances of being wrongly predicted as other expressions. Specifically, *Sadness* is wrongly predicted as *Neutral* by 22.34%, *Disgust* wrongly classified as *Sadness* by 33% and *Contempt* wrongly classified

Table 8
Prediction accuracy of the evaluated models on the test dataset of FER+.

Models	Accuracy
SENet Teacher [2]	88.80%
RAN [41]	89.16%
ViT [8]†	73.36%
FER-VT (ours)†	89.28%
FER-VT (ours)	**90.04%**

*The second-best one.
† Without a pretrained model.



Target label \ Predicted label	Neutral	Happiness	Surprise	Sadness	Anger	Disgust	Fear	Contempt
Neutral	91.48	3.12	1.44	3.36	0.60	0.00	0.00	0.00
Happiness	1.36	96.53	1.36	0.60	0.15	0.00	0.00	0.00
Surprise	4.14	1.91	92.04	0.00	0.96	0.00	0.96	0.00
Sadness	22.34	3.72	2.13	68.62	2.13	0.00	1.06	0.00
Anger	6.63	2.21	4.97	2.21	82.32	0.00	1.66	0.00
Disgust	6.25	6.25	6.25	12.50	6.25	62.50	0.00	0.00
Fear	4.65	0.00	20.93	6.98	0.00	0.00	67.44	0.00
Contempt	50.00	0.00	0.00	12.50	0.00	12.50	0.00	25.00

Fig. 7. Confusion matrix for FER-VT in recognizing eight expressions on FER+ dataset.

as *Neutral* by 50%. Notably, FER-VT still achieves an impressive performance even without a pretrained model on other large-scale facial datasets. In contrast, ViT achieves a just passable performance without a pretrained model. This demonstrates that our token-based visual transformer technique can retain the advantage of VT in learning long-range inductive biases without the requirement of huge facial data for pretraining.

Results on RAF-DB. Table 9 reports the comparisons of FER-VT with other state-of-the-art methods in terms of the overall accuracy and average accuracy on the RAF-DB dataset. For an overall accuracy, FER-VT achieves an accuracy of 88.26%, showing an impressive performance over the other compared models. FER-VT also achieves the best performance over the other compared models in terms of the average accuracy (80.63%). CVT [25] is also a visual transformer-based FER model, but FER-VT achieves a better performance no matter whether it uses a pretrained model or not. This fact suggests that a two-stage A. M. framework can promote the performance of FER models. Fig. 8 shows the confusion matrix for FER-VT (with a pretrained model) on the RAF-DB dataset. FER-VT shows a good capacity of recognizing *Happiness* and *Neutral*, while its accuracy for *Disgust* and *Fear* are only 64.86.62% and 55.62%, respectively. The expressions of *Disgust* were wrongly predicted as *neutral* by 17.50% and the *Disgust* was wrongly predicted as *Disgust* and *Happiness* by 8.12%. As RAF-DB is an in-the-wild dataset, the facial images in the training set are insufficient for ViT to learn the spatial inductive biases as CNN-model models. So, it only achieves an accuracy of 63.75%. While the visual transformer employed in FER-VT has learned the spatial inductive biases from the backbone network, FER-VT still achieves an accuracy of 85.86%, which is close to the performance of state-of-the-art methods. FER-VT makes about a 4.6% improvement over the FER-VT without a pretrained model. Thus, a technique of pretraining a model is also important to the training of a model on a relatively small dataset. Pretraining accelerates the convergence of the loss and promotes the performance of prediction.

4.4. Comparisons on occlusion and pose variant datasets

To show FER-VT is robust to occlusion and variant pose issues on in-the-wild settings, several experiments are conducted on Occlusion-RAF-DB, Pose-RAF-DB, Occlusion-FER+, and Pose-FER+ (these datasets are collected by the work [41] from the FER+ dataset and the RAF-DB dataset). The comparison results on Occlusion-RAF-DB, Pose-RAF-DB, Occlusion-FER+, and Pose-FER+, are reported in Table 10. RAN [41] divides a face image into sub-regions and employs a region biased loss for weighting the importance of different regions for occlusion and pose variant expression recognition. CVT [25] uses a visual transformer technique as FER-VT does. From the table, we can see that both FER-VT and CVT outperform RAN with a large margin on different cases. Specifically, FER-VT outperforms CVT by 0.5% and 0.2%, RAN by 1.9% and 1.16% on Occlusion-FER+ and Occlusion-RAF-DB, respectively. FER also outperforms RAN, CVT on Pose-FER+ and Pose-RAF-DB with pose smaller than 30 degrees. With a pose larger than 45 degrees, FER-VT only outperforms RAN but falls behind CVT, which suggests that the method of division of facial images may not work well to a larger pose degree. Overall, these results have reliably demonstrated that FER-VT is effective in dealing with occlusion and variant pose issues.

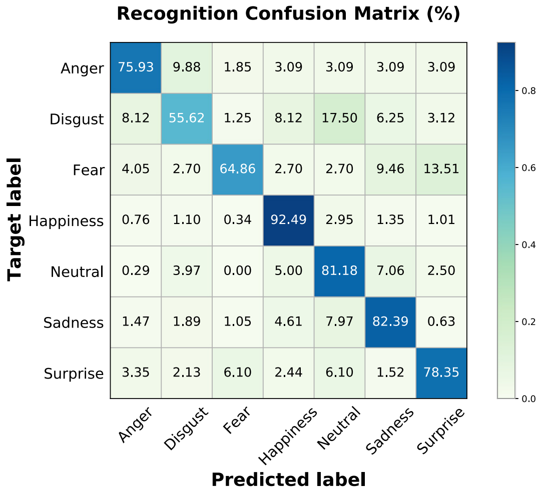
4.5. Visualization and analysis

In this section, we further present our visualization results on investigating the effectiveness of FER-VT with different settings: **1)** we first visualize the feature maps extracted from FER-VT and the pure ResNet to show whether our proposed two-stage mechanism can employ the long-range inductive biases to regulate the feature learning of convolutional filters, and **2)**

Table 9
Recognition performance of the evaluated models on the test set of RAF-DB in terms of overall accuracy and average accuracy.

Models	Overall Acc.	Avg. Acc.
DLP-CNN [20]	84.13%	74.20%
RAN [41]	86.90%	–
DACL [11]†	83.11%	73.20%
DACL [11]	87.78%	80.44%
ViT [8]†	63.75%	–
CVT [25]†	82.27%	–
CVT [25]	88.14%	–
FER-VT(ours)†	84.31%	75.83%
FER-VT(ours)	**88.26%**	**80.63%**

_The second-best performance.
† Without a pretrained model.



	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	75.93	9.88	1.85	3.09	3.09	3.09	3.09
Disgust	8.12	55.62	1.25	8.12	17.50	6.25	3.12
Fear	4.05	2.70	64.86	2.70	2.70	9.46	13.51
Happiness	0.76	1.10	0.34	92.49	2.95	1.35	1.01
Neutral	0.29	3.97	0.00	5.00	81.18	7.06	2.50
Sadness	1.47	1.89	1.05	4.61	7.97	82.39	0.63
Surprise	3.35	2.13	6.10	2.44	6.10	1.52	78.35

Fig. 8. Confusion matrix of recognizing seven expressions on RAF-DB dataset.

we visualize the feature maps extracted from FERT-VT with different grid settings in the GWA neural block. Fig. 9 shows the two facial images that are used to perform the feature visualization.

FER-VT vs **ResNet**. In this visualization, FER-VT with the parameter *grid* 3 × 3 is used to perform the feature extraction. We find that the visualization of the different versions of ResNet will not change the observations and conclusions a lot, to be as concise as possible, we just use ResNet34 as the model of ResNet to depict the observations and conclusions in the following. As shown in Fig. 10, the first column is for the visualization of the feature maps extracted from FER-VT, and the second column is for the visualization of the feature maps extracted from ResNet34. Fig. 10a and b visualize feature maps in the stem layer of FER-VT and ResNet, respectively. It is shown that the contour profiles of the grid-wise attention in FER-VT are more sharpen than those in ResNet, with richer textures. FER-VT can also learn more informative features like LBP [32] than ResNet. The visualizations of feature maps in the second layer of convolutional filter block in FER-VT and ResNet are shown

Table 10
Comparisons on occlusion and pose variant datasets.

Models	Occlusion	Pose(30)	Pose(45)
(a) Results on Occlusion-FER+, Pose-FER+.			
Baseline [41]	73.33%	78.11%	75.50%
RAN [41]	83.63%	82.23%	80.40%
CVT [25]	84.79%	88.29%	**87.20%**
FER-VT (ours)	**85.24%**	**88.56%**	87.06%
(b) Results on Occlusion-RAF-DB, Pose-RAF-DB.			
Baseline [41]	80.19%	84.04%	83.15%
RAN [41]	82.72%	86.74%	85.20%
CVT [25]	83.95%	87.97%	**88.35%**
FER-VT (ours)	**84.32%**	**88.03%**	86.08%

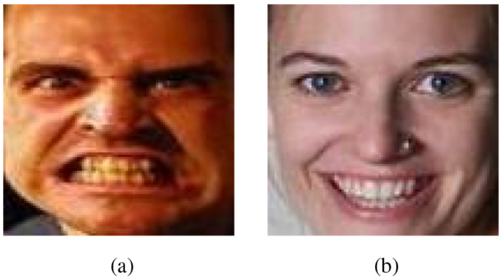
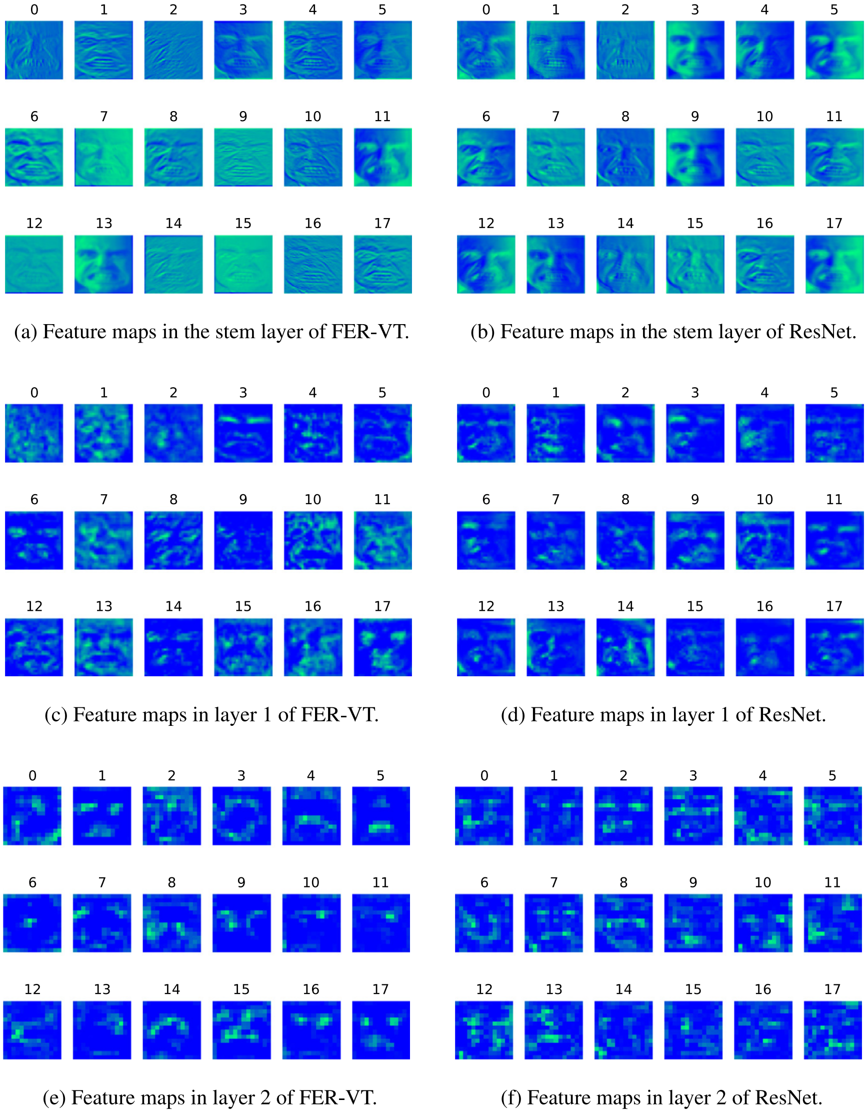

(a) (b)

Fig. 9. The facial images for the visualization of feature maps. (a) The facial image of the expression *Anger* for Fig. 10. (b) The facial image of the expression *Happiness* for Fig. 11.



(a) Feature maps in the stem layer of FER-VT.

(b) Feature maps in the stem layer of ResNet.

(c) Feature maps in layer 1 of FER-VT.

(d) Feature maps in layer 1 of ResNet.

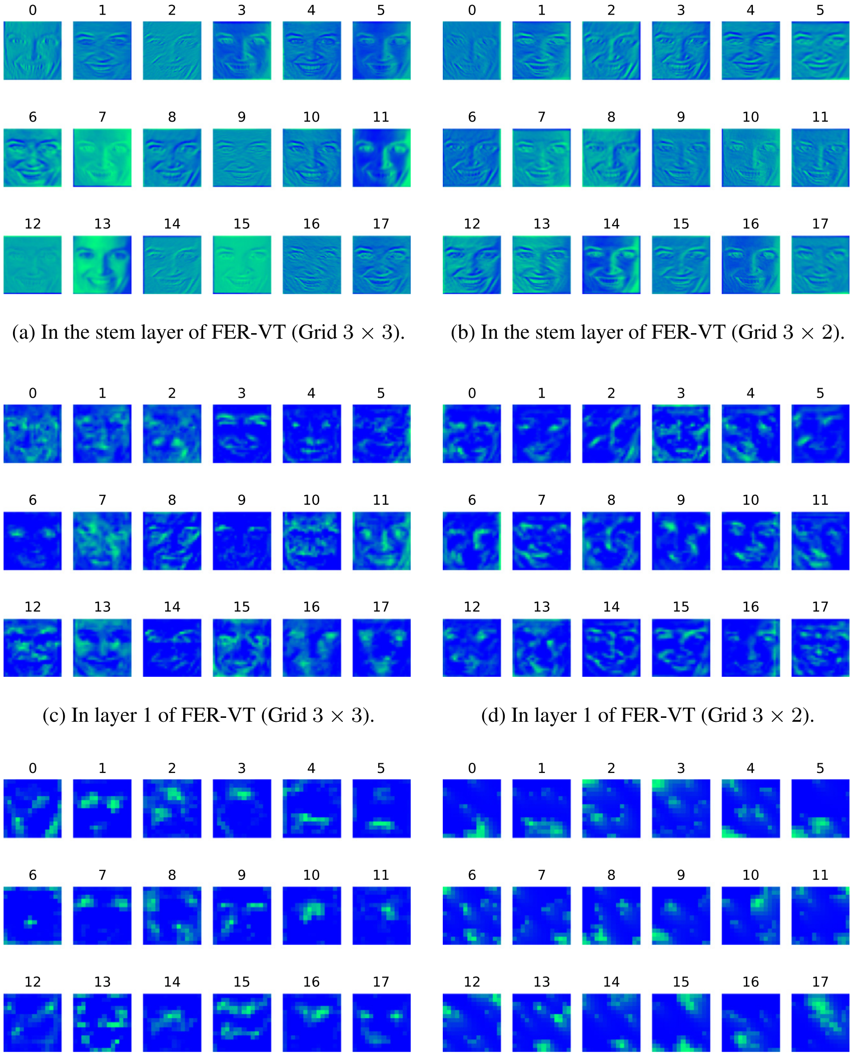
(e) Feature maps in layer 2 of FER-VT.

(f) Feature maps in layer 2 of ResNet.

Fig. 10. Visualization of feature maps from the different layers of FER-VT and ResNet. The visualization figures (i.e. (a), (c), and (e)) in the first column of the figure array are the feature maps from FER-VT. The visualization figures (i.e. (b), (d), and (f)) in the second column of the figure array are the feature maps from ResNet. These feature maps are the first eighteen ones extracted by each layer block.

in Fig. 10c and d, respectively. We can see that FER-VT learns more features from the facial image than ResNet. In addition, FER-VT has a better *understanding* of regions of interest (ROIs) than ResNet, as more features are learned from these ROIs. Fig. 10e and f are for feature maps in the third layer of convolutional filter block in FER-VT and ResNet, respectively. These feature maps are extracted from the deeper convolutional filter blocks of FER-VT and ResNet, resulting in more abstract features. We can see that feature maps in FER-VT are more semantic and meaningful for FER, while the features in ResNet are more stochastic with much *noise*. Thus, with the long-range inductive biases learned from *GWA* and *VTA*, the parameters learning of convolutional filters in FER-VT can be regularized in a way that the important facial units for FER are focused on while reducing the *noise* of features as much as possible.

Grid 3 × 3 vs **Grid** 3 × 2. We visualize the feature maps extracted from the FER-VT with the *grid* parameter as 3 × 3 and the FER-VT with the *grid* parameter as 3 × 2. As shown in Fig. 11, the first column is for the visualization of the feature maps extracted from the FER-VT with the *grid* parameter as 3 × 3, and the second column is for the visualization of the feature maps extracted from the FER-VT with the *grid* parameter as 3 × 3. Fig. 11a and b are the visualizations of feature maps in the stem layer of the FER-VT (*grid* 3 × 3) and the FER-VT (*grid* 3 × 2), respectively. It is shown that both the models show a similar pattern to extract features from the facial image, while the FER-VT (*grid* 3 × 3) trends to keep more features that are similar to the original facial image than the FER-VT(*grid* 3 × 2). Fig. 11c and d are the visualization of feature maps in the



(a) In the stem layer of FER-VT (Grid 3 × 3).

(b) In the stem layer of FER-VT (Grid 3 × 2).

(c) In layer 1 of FER-VT (Grid 3 × 3).

(d) In layer 1 of FER-VT (Grid 3 × 2).

(e) In layer 2 of FER-VT (Grid 3 × 3).

(f) In layer 2 of FER-VT (Grid 3 × 2).

Fig. 11. Visualization of feature maps from the different layers of the FER-VT (Grid 3 × 3) and the FER-VT (Grid 3 × 2). The visualization figures (i.e. (a), (c), and (e)) in the first column of the figure array are the feature maps from the FER-VT (Grid 3 × 3). The visualization figures (i.e. (b), (d), and (f)) in the second column of the figure array are the feature maps from the FER-VT (Grid 3 × 2). These feature maps are the first eighteen ones extracted by each layer block.

second layer of convolutional filter block in the FER-VT (*grid* 3 × 3) and the FER-VT (*grid* 3 × 2), respectively. We can see that the FER-VT (*grid* 3 × 3) learns more features from the facial image than the FER-VT (*grid* 3 × 2). Fig. 11e and f are the visualizations of feature maps in the third layer of convolutional filter blocks in the FER-VT (*grid* 3 × 3) and the FER-VT(*grid* 3 × 2), respectively. We can see that feature maps in the FER-VT(*grid* 3 × 3) are easier to be recognized with human intuition. We also conducted an experiment on the FER + dataset with the different settings of the parameter *grid*, e.g. (*grid* 3 × 2), (*grid* 3 × 3) and (*grid* 4 × 4). The (*grid* 3 × 3) performs the best performance, followed by the (*grid* 4 × 4), and the (*grid* 3 × 2) is the last. Thus, the parameter *grid* is an important factor in improving the performance of CNN-based models for FER. We will investigate this setting more in future work.

4.6. Ablation study

We conduct an extensive ablation study to demonstrate the effectiveness of different components of our proposed model FER-VT and different alternative blocks in GWA. The results by using different strategies in pyramid feature fusion are also studied.

Ablation study on individual components. We first examine the contributions of individual components in FER-VT. As shown in Table 11, the ResNet34 backbone achieves the accuracy of 95.60%, 86.74%, and 76.83% on three datasets: CK+, FER+, and RAF-DB, respectively, which may be different from some existing methods because of our designed training process. With the grid-wise attention mechanism, the *ResNet+GWA* outperforms the backbone by 4.01%, 2.66%, and 9.61% on three datasets respectively. In particular, the improvement in the RAF-DB dataset is very impressive. In Section 4.5, the visualization has shown that the grid-wise attention mechanism helps to guide convolutional filters in the stem layer to learn richer features on facial contour profile and textures for the deeper convolutional filters for FER. The combination of *ResNet+VTA* performs a better performance than the combination of *ResNet + GWA*, outperforming the backbone by 4.16%, 2.81% and 10.18% on three datasets respectively. The improvement on the RAF-DB is above 10% higher over compared to the backbone network. In Section 4.5, the visualization has also shown that the VTA mechanism helps convolutional filters in the deeper layer to focus on the important facial units or ROIs and reduce the possible *noisy* features for FER. The *ResNet+GWA+VTA* obtains the best performance in these combinations and outperforms the backbone by 4.35%, 2.93%, and 10.78% on three datasets respectively. The results have shown that both these two components are effective in CNN-based models for FER. The performance of the *ResNet+GWA+VTA* has demonstrated that the two-stage attention mechanism performs better than either a one-stage attention mechanism or two-stage attention with an identical mechanism.

Effectiveness on alternative blocks for local feature extraction. In this section, we study the effects on the alternative block for local feature extraction presented in Section 3.3.1. As shown in Table 12, the inverted bottleneck block (IBN block) used in this paper achieves the best performance, with an accuracy of 100% on the CK + dataset, 89.28% on the FER+ dataset, and 84.31% on the RAF-DB dataset, followed by the local relation network block (LRN block) (99.87% on CK+, 89.25% on FER+, and 84.72% on RAF-DB). The idea of the IBN block is inspired by the work in MobileNetV2 [31], which also points that the block preserves more information for the original input than the BN block [14], and the reported results also show the IBN block performs better. The LRN block is a bottom-up feature extraction neural block for the work [16]. This block can achieve a more adaptive feature extraction than convolutional filters due to a bottom-up method, as the parameters in convolutional filters are static after training [16]. This block achieves better performance over the BN block, but still not outperforms the IBN block.

Effectiveness on alternative pyramid feature fusion strategies. The token-based visual transformer is a fusion strategy for pyramid features indeed. In this section, we study the effects of alternative pyramid feature fusion strategies. As shown in Table 13, both the ASSF and VTA (ours) achieve better results than the one without using feature fusion strategies (FC) (97.80% for CK+, 89.05% for FER+, and 84.31% for RAF-DB). It shows that the feature fusion strategies in the deeper convolutional filters can improve the performance of CNN-based Models for FER. The ASSF strategy employs a learnable weight vector to quantify the importance of features from the different convolutional filter blocks. This is also a kind of attention mechanism that can learn long-range inductive biases in these feature maps. But the VTA uses a transformer as a feature fusion network. This transformer has a stronger capacity for learning long-range inductive biases than the simple attention mechanism in ASSF. Thus, the VTA achieves a better performance than the ASSF.

Table 11
Ablation study on the individual components.

Combinations	CK+	FER+	RAF-DB
ResNet†	95.60%	86.74%	76.83%
ResNet + GWA†	97.80%	89.05%	84.31%
ResNet + VTA†	98.90%	89.18%	84.65%
ResNet + GWA + VTA†	**100.00%**	**89.28%**	**84.31%**

*The second-best performance.
† Without a pretrained model.

Table 12
The effectiveness of alternative neural block for the local feature extraction in GWA.

Alternative Blocks	CK+	FER+	RAF-DB
BN Block[†,1]	99.79%	89.12%	84.36%
LRN BLOCK[†,2]	99.87%	89.25%	84.72%
IBN Block (Ours)[†,3]	**100%**	**89.28%**	**84.31%**

*The second-best performance.
[1] A bottleneck neural block from ResNet [14].
[2] A bottom-up feature extraction block from [16].
[3] The inverted bottleneck neural in Section 3.3.1.
[†] Without a pretrained model.

Table 13
The effectiveness of alternative pyramid feature fusion strategies for FER task.

Strategies	CK+	FER+	RAF-DB
FC[†,1]	97.80%	89.05%	84.31%
ASSF[†,2]	99.79%	89.14%	84.37%
VTA (Ours)[†,3]	**100**%	**89.28**%	**84.31**%

*The second-best performance.
[1] Without a fusion strategy.
[2] The fusion strategy from [22].
[3] The proposed attention strategy in Section 3.4.
[†] Without a pretrained model.

5. Conclusion

To overcome the weakness of CNN-based FER models in learning long-range inductive biases, this paper introduces a novel framework of a two-stage attention mechanism named FER-VT for CNN-based FER models to improve its performance on a FER task. In particular, the two attention mechanisms are used at the low-level feature learning stage and the high semantic representation stage, respectively. In the first stage, grid-wise attention (GWA) captures the dependencies among different regions from a facial expression image in a way that the parameter learning of convolutional filters is regularized. At the high semantic representation stage, a token-based Visual Transformer (VT) learns the global representation from a sequence of visual semantic tokens. To the best of our knowledge, it is the first work on employing a token-based VT technique for a FER task. The results of our extensive experiments have shown that the two-stage A.M. FER model of FER-VT outperforms state-of-the-art A.M.-based FER models in terms of accuracy. This implies that a two-stage A.M. framework can promote the performance of CNN-based FER models.

Our future work will perform an attention graph neural network (GNN) in the low-level feature learning stage. This is because GNN may have a stronger capacity for modeling the long-range dependencies between different facial units of a facial expression image.

CRediT authorship contribution statement

Qionghao Huang: Conceptualization, Methodology, Software, Writing - original draft. **Changqin Huang:** Supervision, Formal analysis, Writing - review & editing. **Xizhe Wang:** Investigation, Visualization. **Fan Jiang:** Data curation, Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant (Nos. 61877020, 62007031), the Science and Technology Project of Guangzhou Municipality, China under Grant (No. 201904010393), and the Key Research and Development Program of Zhejiang Province (No. 2021C03141).

References

[1] A. Agrawal, N. Mittal, Using CNN for facial expression recognition: A study of the effects of kernel size and number of filters on accuracy, The Visual Computer 36 (2) (2020) 405–412.

[2] S. Albanie, A. Nagrani, A. Vedaldi, A. Zisserman, Emotion recognition in speech using cross-modal transfer in the wild, in: Proceedings of the 26th ACM International Conference on Multimedia, ACM, 2018, pp. 292–301.

[3] S.A. Bargal, E. Barsoum, C.C. Ferrer, C. Zhang, Emotion recognition in the wild from videos using images, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, ACM, 2016, pp. 433–436.

[4] E. Barsoum, C. Zhang, C.C. Ferrer, Z. Zhang, Training deep networks for facial expression recognition with crowd-sourced label distribution, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, ACM, 2016, pp. 279–283.

[5] J.J. Bazzo, M.V. Lamar, Recognizing facial actions using gabor wavelets with neutral face average difference, in: Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition, IEEE, 2004, pp. 505–510.

[6] J. Chen, R. Xu, L. Liu, Deep peak-neutral difference feature for facial expression recognition, Multimedia Tools and Applications 77 (22) (2018) 29871–29887.

[7] Y. Chen, J. Wang, S. Chen, Z. Shi, J. Cai, Facial motion prior networks for facial expression recognition, in: Proceedings of the 2019 IEEE Visual Communications and Image Processing, IEEE, 2019, pp. 1–4.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al, An image is worth 16×16 words: Transformers for image recognition at scale, in: Proceedings of the 2021 International Conference on Learning Representations, OpenReview.net, 2020, pp. 1–21.

[9] L. Du, H. Hu, Modified classification and regression tree for facial expression recognition with using difference expression images, Electronics Letters 53 (9) (2017) 590–592.

[10] Y. Fan, J.C. Lam, V.O. Li, Video-based emotion recognition using deeply-supervised neural networks, in: Proceedings of the 2018 International Conference on Multimodal Interaction, ACM, 2018, pp. 584–588.

[11] A.H. Farzaneh, X. Qi, Facial expression recognition in the wild via deep attentive center loss, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, IEEE, 2021, pp. 2402–2411. .

[12] I.J. Goodfellow, D. Erhan, P.L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al, Challenges in representation learning: A report on three machine learning contests, in: Proceedings of the 2013 International Conference on Neural Information Processing, Springer, 2013, pp. 117–124.

[13] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al., A survey on visual transformer, arXiv preprint arXiv:2012.12556. .

[14] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 770–778.

[15] M.S. Hossain, G. Muhammad, Emotion recognition using secure edge and cloud computing, Information Sciences 504 (2019) 589–601.

[16] H. Hu, Z. Zhang, Z. Xie, S. Lin, Local relation networks for image recognition, in: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, IEEE, 2019, pp. 3464–3473.

[17] H. Jun, L. Shuai, S. Jinming, L. Yue, W. Jingwei, J. Peng, Facial expression recognition based on VGGNet convolutional neural network, in: Proceedings of the 2018 Chinese Automation Congress, IEEE, 2018, pp. 4146–4151.

[18] S. Khan, M. Naseer, M. Hayat, S.W. Zamir, F.S. Khan, M. Shah, Transformers in vision: A survey, arXiv preprint arXiv:2101.01169. .

[19] P. Kumar, P.P. Roy, D.P. Dogra, Independent bayesian classifier combination based sign language recognition using facial expression, Information Sciences 428 (2018) 30–48.

[20] S. Li, W. Deng, Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition, IEEE Transactions on Image Processing 28 (1) (2018) 356–370.

[21] S. Li, W. Deng, Deep facial expression recognition: A survey, IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2020.2981446. .

[22] S. Liu, D. Huang, Y. Wang, Learning spatial fusion for single-shot object detection, arXiv preprint arXiv:1911.09516. .

[23] Y. Liu, J. Peng, J. Zeng, S. Shan, Pose-adaptive hierarchical attention network for facial expression recognition, arXiv preprint arXiv:1905.10059. .

[24] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression, in: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2010, pp. 94–101.

[25] F. Ma, B. Sun, S. Li, Robust facial expression recognition with convolutional visual transformers, arXiv preprint arXiv:2103.16854. .

[26] D. Meng, X. Peng, K. Wang, Y. Qiao, Frame attention networks for facial expression recognition in videos, in: Proceedings of the 2019 IEEE International Conference on Image Processing, IEEE, 2019, pp. 3866–3870.

[27] S. Minaee, M. Minaei, A. Abdolrashidi, Deep-emotion: Facial expression recognition using attentional convolutional network, Sensors 21 (9) (2021) 3046.

[28] K. Mohan, A. Seal, O. Krejcar, A. Yazidi, Facial expression recognition using local gravitational force descriptor-based deep convolution neural networks, IEEE Transactions on Instrumentation and Measurement 70 (2020) 1–12.

[29] M. Pantic, M. Valstar, R. Rademaker, L. Maat, Web-based database for facial expression analysis, in: Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, IEEE, 2005, pp. 5–9.

[30] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, J. Shlens, Stand-alone self-attention in vision models, Proceedings of the 33rd Conference on Neural Information Processing Systems, vol. 32, Springer, 2019, pp. 1–13.

[31] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.C. Chen, Mobilenetv 2: Inverted residuals and linear bottlenecks, in: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2018, pp. 4510–4520.

[32] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: A comprehensive study, Image and Vision Computing 27 (6) (2009) 803–816.

[33] J. Shao, Y. Qian, Three convolutional neural network models for facial expression recognition in the wild, Neurocomputing 355 (2019) 82–92.

[34] F. Sultana, A. Sufian, P. Dutta, Evolution of image segmentation using deep convolutional neural network: A survey, Knowledge-Based Systems 201 (2020) 106062.

[35] X. Sun, P. Xia, L. Zhang, L. Shao, A ROI-guide deep architecture for robust facial expressions recognition, Information Sciences 522 (2020) 35–48.

[36] M. Tan, R. Pang, Q.V. Le, Efficientdet: Scalable and efficient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10781–10790.

[37] Y.I. Tian, T. Kanade, J.F. Cohn, Recognizing action units for facial expression analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (2) (2001) 97–115.

[38] M. Tkalčič, A. Odić, A. Košir, The impact of weak ground truth and facial expressiveness on affect detection accuracy from time-continuous videos of facial expressions, Information Sciences 249 (2013) 13–23.

[39] Y. Tong, W. Liao, Q. Ji, Facial action unit recognition by exploiting their dynamic and semantic relationships, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (10) (2007) 1683–1699.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the Advances in Neural Information Processing Systems, vol. 30, Springer, 2017, pp. 5998–6008. .

[41] K. Wang, X. Peng, J. Yang, D. Meng, Y. Qiao, Region attention networks for pose and occlusion robust facial expression recognition, IEEE Transactions on Image Processing 29 (2020) 4057–4069.

[42] X. Wang, P. Wu, G. Liu, Q. Huang, X. Hu, H. Xu, Learning performance prediction via convolutional GRU and explainable neural networks in e-learning environments, Computing 101 (6) (2019) 587–604.

[43] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, M. Tomizuka, K. Keutzer, P. Vajda, Visual transformers: Token-based image representation and processing for computer vision, arXiv preprint arXiv:2006.03677. .

[44] K. Zhang, Y. Huang, Y. Du, L. Wang, Facial expression recognition based on deep evolutional spatial-temporal networks, IEEE Transactions on Image Processing 9 (9) (2017) 4193–4203.

[45] G. Zhao, X. Huang, M. Taini, S.Z. Li, M. PietikäInen, Facial expression recognition from near-infrared videos, Image and Vision Computing 29 (9) (2011) 607–619.

[46] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, S. Yan, Peak-piloted deep network for facial expression recognition, Proceedings of the 2016 European Conference on Computer Vision, vol. 9906, Springer, 2016, pp. 425–442.

[47] H. Zheng, R. Wang, W. Ji, M. Zong, W.K. Wong, Z. Lai, H. Lv, Discriminative deep multi-task learning for facial expression recognition, Information Sciences 533 (2020) 60–71.

[48] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, D.N. Metaxas, Learning active facial patches for expression analysis, in: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2562–2569.