Robust Facial Expression Recognition with Convolutional Visual Transformers

Fuyan Ma, Bin Sun, *Member, IEEE,* and Shutao Li, *Fellow, IEEE*

Abstract—Facial Expression Recognition (FER) in the wild is extremely challenging due to occlusions, variant head poses, face deformation and motion blur under unconstrained conditions. Although substantial progresses have been made in automatic FER in the past few decades, previous studies were mainly designed for lab-controlled FER. Real-world occlusions, variant head poses and other issues definitely increase the difficulty of FER on account of these information-deficient regions and complex backgrounds. Different from previous pure CNNs based methods, we argue that it is feasible and practical to translate facial images into sequences of visual words and perform expression recognition from a global perspective. Therefore, we propose Convolutional Visual Transformers to tackle FER in the wild by two main steps. First, we propose the attentional selective fusion (ASF) for leveraging two kinds of feature maps generated by two-branch CNNs. The ASF captures discriminative information by fusing multiple features with the global-local attention. The fused feature maps are then flattened and projected into sequences of visual words. Second, inspired by the success of Transformers in natural language processing, we propose to model relationships between these visual words with the global self-attention. The proposed method is evaluated on three public in-the-wild facial expression datasets (RAF-DB, FERPlus and AffectNet). Under the same settings, extensive experiments demonstrate that our method shows superior performance over other methods, setting new state of the art on RAF-DB with 88.14%, FERPlus with 88.81% and AffectNet with 61.85%. The cross-dataset evaluation on CK+ shows the promising generalization capability of the proposed method.

Index Terms—Facial expression recognition in the wild, global-local attention, Transformers, global self-attention.

1 Introduction

Understanding human emotional states is the fundamental task to develop emotional intelligence, which is an interdisciplinary field spanning from different research areas, such as psychology and computer science. Facial expression is one of the most natural, powerful and universal signals for human beings to convey their emotional states and intentions [1], [2]. Facial expression recognition (FER) systems have various applications, including human-robot interaction (HRI), mental health assessment and driver fatigue monitoring. Therefore, numerous research endeavours have been invested for promoting the development of FER.

As shown in Fig. 1, the challenges of FER in the wild mainly come from occlusions, variant head poses, face deformation, motion blur, insufficient qualitative data, etc., which lead to significant changes of the facial appearance. These unexpected issues definitely increase the difficulty of expression recognition. With the advance of machine learning, especially deep learning, researchers have made great progress on FER in the past decades. Before deep learning era, traditional FER methods have mainly used handcrafted features and shallow learning (e.g., Histograms of Oriented Gradients (HOGs) [3], Local Binary Patterns (LBP) [4], [5], non-negative matrix factorization (NMF) [6] and sparse representation [7]). With the popularity of data-driven techniques, deep learning based methods have exceeded traditional methods by a large margin and achieved state-of-the-art FER performance (e.g., [8], [9], [10], [11]).

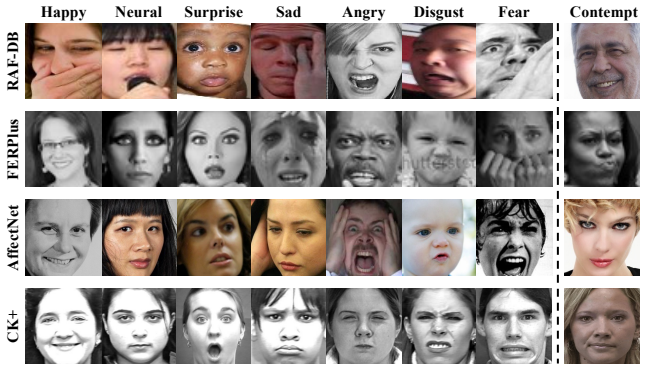


Fig. 1. The samples from RAF-DB, FERPlus, AffectNet and CK+. Variant head poses, occlusions and other unconstrained conditions can be seen in above images. Note that RAF-DB is annotated with seven basic expressions and FERPlus, AffectNet, CK+ are with eight expression labels, including the contempt category.

Despite the significant success of learned representation for constrained FER, the performance of FER in the wild is still far from satisfactory. Majority of these proposed algorithms are implemented on lab-collected datasets, such as CK+ [12], MMI [13] and Oulu-CASIA [14]. These algorithms perform perfectly on these lab-collected FER datasets, be-

- *This work is supported by the National Key Research and Development Project (2018YFB1305200), the National Natural Science Fund of China (61801178) and the Key-Area Research and Development Plan of Guangdong Province (2018B010107001).*
- *Fuyan Ma and Bin Sun are with College of Electrical and Information Engineering, and with the Key Laboratory of Visual Perception and Artificial Intelligence of Hunan Province, Hunan University, Changsha, 410082, China. (mafuyan@hnu.edu.cn; sunbin611@hnu.edu.cn)*
- *Shutao Li is with College of Electrical and Information Engineering, with the State Key Laboratory of Advanced Design and Manufacturing for Vehicle Body and with the Key Laboratory of Visual Perception and Artificial Intelligence of Hunan Province, Hunan University, Changsha, 410082,China.(shutao_li@hnu.edu.cn)*

cause the controlled images are frontal with minimal illumination changes and limited occlusions. However, the performance degrades dramatically on the real-world FER datasets, such as RAF-DB [8], FERPlus [15] and AffectNet [16]. Compared with the real-world FER datasets, the number of images from the lab-collected FER datasets is relatively small. Especially, convolutional neural networks (CNNs) [17] for the FER in the wild usually require sufficient training face images to ensure generalizability for real applications. Most publicly available datasets for FER do not have a sufficient quantity of images for training. Therefore, the performance of FER in the wild is limited by not only the unconstrained conditions but also the available data volume.

Different backgrounds, illuminations and head poses are fairly common under the unconstrained conditions, which are irrelevant to facial expressions [18]. Directly recognizing expression on such images is a big challenge. To remove complex backgrounds and non-face areas, it is indispensable to detect faces before training the deep neural network to learn meaningful features. Several face detectors like MTCNN [19] and Dlib [20] are used to detect faces in complex scenarios. The face alignment process is also crucial because it can reduce the variation in facial size and in-plane rotation. After obtaining relatively accurate face regions, various methods have been proposed to improve the FER accuracy and enhance the generalization ability of expression recognition algorithms. Most of the researchers design their methods for occlusion-aware and multi-view FER, which are two main obstacles for FER in real-world scenarios. Bourel et al. [21] proposed to recover facial feature points for the recognition of facial expressions in the presence of occlusion. The appearance features of salient facial patches have also been used for removing features of occlusion patches in [22]. More recently, attention models have been successfully applied for FER to explore meaningful regions. Li et al. [23] proposed a patch-gated CNN that integrates path-level attention for expression recognition with occlusion. Besides, attention models have been successfully applied for FER to explore meaningful regions. Similar to [23], several methods, such as [10], [24], [25], also used attention-like mechanisms to focus on the most discriminative features to improve FER accuracy. Non-frontal face images are overwhelmingly common in real-world scenarios. Previous methods often treat expression recognition in non-frontal face images as the multi-view FER problem. Zheng [26] proposed to select the optimal sub-regions of a face that contribute most to the expression recognition based on a group sparse reduced-rank regression. Liu et al. [27] proposed to tackle multi-view FER by three parts: the multi-channel feature extraction, the multi-scale feature fusion and the pose-aware recognition. In addition, generative adversarial networks (GAN) have also been applied for multi-view FER. GAN-based methods [28] [29] can synthesize face images with large head pose variations to enlarge the training set for FER. Especially, Zhang et al. [30] proposed an end to end model based on GAN for facial images synthesize with a set of facial landmarks and pose-invariant facial expression recognition by exploiting the geometry information. Sun et al. [31] proposed cyclic image generation for unsupervised cross-view facial expression recognition based on GAN.

Taking variant poses and occlusions in face images for example, we illustrate the motivation of our method. Variant poses and the resize operation lead to face deformation compared with frontal faces, which can be seen as out-of-order subregions. Occlusions in face images result in information-deficient subregions. By analogy with natural language, FER in the wild can be tackled in a different way. There is an interesting phenomenon that we can understand the sentences, i.e., "*I cdn'uolt blveiee taht I cluod aulaclty undresatnd what I was rdanieg: the phaonmneel pweor of the hmuan mnid.*" One of the most possible reasons is that the cognitive mode of us human beings usually impels ourselves to look and think about this sentence globally. Also like the image description task in Natural Language Processing (NLP), we may use several visual words to describe the emotional state for an image. Inspired by this cognitive mode, we hypothesize that it is feasible and effective to recognize the facial expressions by a sequence of visual words from a global perspective. Therefore, we propose Convolutional Visual Transformers (CVT) for robust facial expression recognition in the wild. Our method combines LBP features and CNN features to further enrich the representation of the visual words, referring to the hybrid feature extraction. The reason we use LBP features is that it can catch the small movements of the faces and extract image texture information. We design the attentional feature fusion (ASF) to adaptively integrate LBP features and CNN features. The ASF aggregates both global and local relationships between two kinds of features, which can effectively improve the recognition performance. We simply convert the fused feature maps into a sequence of visual words by flattening and projecting the features maps. After obtaining these visual words, we then exploit the multi-layer Transformer encoder to boost the performance. The global self-attention in the multi-layer Transformer encoder allows the network to model the contextual information of the representative visual words and focus on the most discriminative features.

Overall, the main contributions of our work can be summarized as follows:

1) We propose Convolutional Visual Transformers for FER in the wild, which integrates LBP features and CNN features with the global-local attention and the global self-attention for improving expression recognition accuracy.
2) We design a simple but effective feature fusion module named ASF to aggregate both global and local facial information. Moreover, the ASF guides the backbones to extract the required information while squeeze the useless information in an end-to-end manner.
3) To the best of our knowledge, we are the first to apply Transformers for the FER. The global self-attention enables the whole network to learn the relationships between elements of visual feature sequences and ignore the information-deficient regions.
4) Extensive experimental results on three publicly available FER in the wild datasets, i.e., RAF-DB, FERPlus and AffectNet, demonstrate that our CVT

achieves state-of-the-art expression recognition performance. Especially, we also conduct experiments on occlusion and variant pose subsets of these three datasets and cross-dataset evaluation on CK+ to show the promising generalization ability of our method.

The rest of this article is organized as follows: Section 2 briefly reviews related works for FER in the wild and provides a comprehensive review of recent advances in FER. The core idea of our proposed CVT is presented in Section 3. Section 4 presents an extensive performance evaluation for the proposed method and state-of-the-art approaches. We conclude our work in Section 5.

2 Related Work

2.1 Facial Expression Recognition in the Wild

Facial expression recognition has been an emerging topic with the development of deep learning, especially convolutional neural networks. The CNNs have shown good results for FER on lab-collected datasets. However, the performance of FER in the wild is heavily influenced by unexpected conditions mentioned in Section 1. The critical step for FER in the wild is to accurately extract discriminative features. Researchers start to shift their attention from hand-crafted features to deep features for facial expression recognition. Various architectures have been proposed for extracting deep features, such as Deep Belief Networks, ResNet [32], DenseNet [33]. Tang [34] utilized the CNNs for feature extraction and replaced the softmax layer with the linear SVMs, which gave significant gains and won the ICML 2013 Representation Learning Workshop's face expression recognition challenge.

To recognize facial expressions in the wild, Li et al. [8] proposed to enhance the discriminative power of deep features by a deep locality-preserving CNN (DLP-CNN) method. Region-based attention networks are especially suitable for FER in the wild, because they allow for salient face features to dynamically come to forefront when some occlusions or clutters occur in an image. In [10], Wang et al. proposed region attention networks (RAN) to capture the importance of facial regions for occlusion and pose variant FER. Likewise, Li [9] et al. proposed a CNN with attention mechanism for Occlusion Aware FER, which focused on the most discriminative face regions. Some FER algorithms integrated demographic features (i.e., gender, race, age, etc.) for improving the expression recognition performance. It is worth mentioning that Fan [24] proposed a deeply-supervised attention network, which takes facial attributes into consideration. In addition, Xu et al. [35] conducted a comparative study of the bias and fairness for FER and use the attribute information as input to address bias. Wang et al. [11] proposed a self-cure Network (SCN) to suppress the uncertainties for FER in the wild. SCN also took full advantage of attention mechanism to weight each training face sample. To improve FER accuracy in the wild, we incorporate the global-local attention to integrate both LBP features and CNN features to get informative visual words. We also apply the global self-attention to model relationships between visual words, which allows for suppressing confusing regions and highlighting discriminative features.

2.2 Feature Fusion for FER

Multi-scale feature fusion in deep learning has been widely invested for a multitude of applications, such as face recognition [36], object detection [37], semantic segmentation [38]. However, deep learning based methods require huge amount of data to generalize well. Fusing different feature maps can enrich the representative ability of the whole networks, which can effectively improve the generalization ability and the recognition performance. As we mentioned in Section 1, there are various hand-crafted features, such as HOGs [39] and LBP [40], which have been well-developed for facial expression feature representation. Besides, CNNs have been proposed for FER in the wild due to the powerful representation capability. Chen et al. [41] proposed to fuse dynamic textures, geometric features and acoustic features to tackle FER in the wild. The dynamic texture descriptor of visual information is an extension of HOGs, named Histogram of Oriented Gradients from Three Orthogonal Planes (HOG-TOP). Shao and Qian [42] proposed a dual branch CNN to extracts LBP features and deep features parallel. The concatenation operation was then utilized to fuse these feature maps. Li et al. [25] combined LBP features and CNN features with dense connection to improve expression recognition accuracy. Previous feature fusion method often used element-wise summation or concatenation operation as the fusion strategy. They did not well utilize complementary information and abandoned redundant information. Additionally, these feature fusion methods were mainly conducted on lab-controlled FER datasets, such as CK+ and Oulu-CASIA. Different from existing methods, we propose the attentional selective fusion (ASF) for integrating the LBP features and the CNN features, which squeezes the useless information and generates correlated weight maps from both local and global perspectives. The ASF is applied for FER datasets in the wild (RAF-DB, FERPlus and AffectNet). The results in Table 5 prove the effectiveness of our ASF.

2.3 Transformers in Computer Vision

Transformers [43] have shown dominant performance in NLP. Inspired by the success of Transformers, several researchers have tried to invest Transformers on computer vision tasks, such as object detection [44], pose estimation [45], high-resolution image synthesis [46], video instance segmentation [47], trajectory prediction [48], etc. Vision Transformer (ViT) [49] was the first work to apply a vanilla Transformer to images with few modifications. ViT directly split an image into patches and fed these patches into a Transformer. According to [49], ViT yielded lower accuracy compared with ResNet when trained on ImageNet [50]. ViT was firstly trained on large datasets, and then fine-tuned for downstream tasks, because Transformers need amounts of data to generalize well on computer vision tasks. Wang et al. [51] proposed Pyramid Vision Transformer (PVT) for pixel-level dense prediction. PVT can work as the feature extraction backbone without convolutions, but with feature pyramid structure. Both ViT and PVT are pure Transformers without convolution operation. Transformer-based approaches have shown superior performance compared with CNN-based methods, when fully trained on large-scale datasets. Inspired by the vanilla Transformer and ViT,

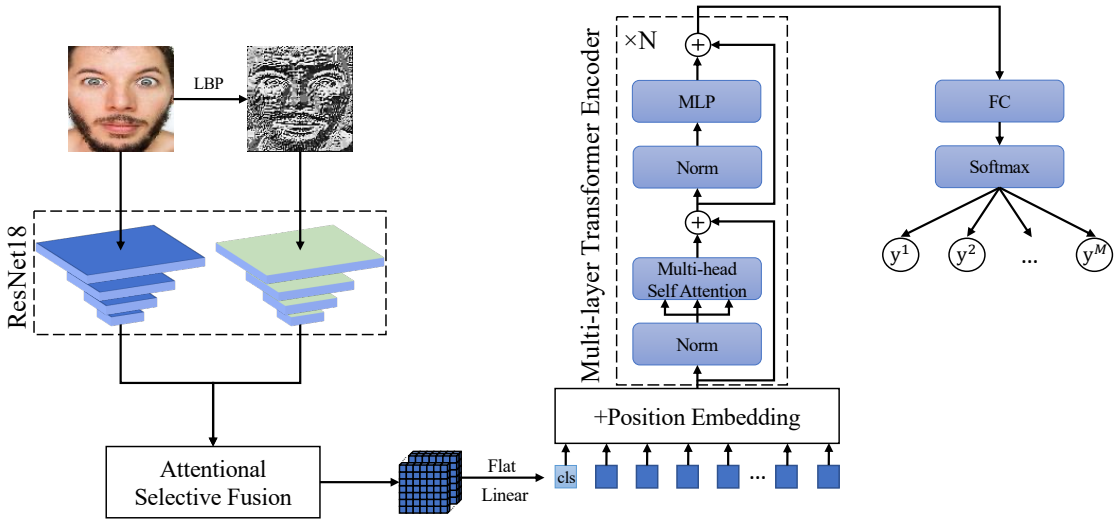


Fig. 2. An overview of our proposed Convolutional Visual Transformers. It can be divided into three parts, visual words extraction, relationship modeling and expression classification. The pre-trained ResNet18 is used as the backbone to extract feature maps. All the extracted features are fused by our attentional selective fusion to get representative visual words. The input visual words are obtained by simply flattening the spatial dimensions of the feature maps and projecting to the specific dimension. And then apply the multi-layer Transformer encoder to model the relationships between different visual features components. The network finally calculates the expression probabilities by a simple softmax function.

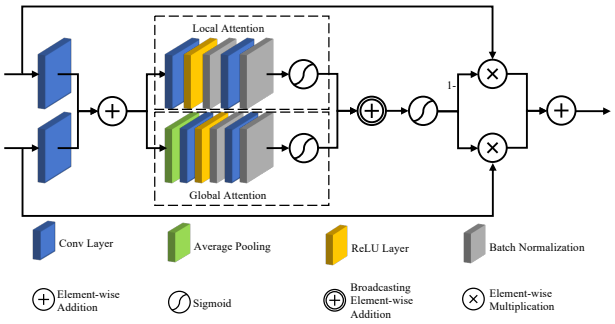


Fig. 3. The attentional selective fusion module.

we firstly propose to directly apply Transformers for FER. The Transformers model long dependencies between input sequences by using the global self-attention mechanism. Such global self-attention enables the model to ignore the information-deficient regions and recognize the expressions from a global perspective in case of occlusions or variant poses for FER.

3 Method

In this section, we first elaborate the overview of the proposed Convolutional Visual Transformers for facial expression recognition. Then, we explicitly illustrate the attentional selective fusion, and the multi-Layer Transformer encoder in our whole network.

3.1 Overview

Fig. 2 illustrates the overall diagram of our Convolutional Visual Transformers for facial expression recognition. Our CVT is built upon on two pre-trained ResNet18 [32] networks, and consists of two crucial components: i) attentional selective fusion, ii) multi-Layer Transformer encoder.

For a given face image I_{RGB} with the size of $H \times W \times 3$, we first get its LBP feature image with the size of $H \times W \times 1$ and concatenate it to a feature image I_{LBP} with the size of $H \times W \times 3$. The feature extraction backbones are composed of two ResNet18 networks: one is for the RGB image and the other is for its LBP feature image. Particularly, we employ the first five stages of ResNet18 as the backbone to extract feature maps X_{LBP} and X_{RGB} with the size of $\frac{H}{R} \times \frac{W}{R} \times C_f$, where R is the downsampling rate of ResNet18, C_f is the channel number of the output of the stage 5. For simplicity, we denote $H_d = \frac{H}{R}$ and $W_d = \frac{W}{R}$. In this paper, $R = 32$ and $H = W = 224$. We initialize the whole network weights by the pre-trained weights on MS-Celeb-1M face recognition dataset. Without loss of generalization, the ASF is utilized to combine the features extracted from the RGB image and the features extracted from its LBP feature image, which will be introduced in Section 3.2 in detail. The ASF module dynamically adjusts the weights of these features and guides the networks focus more on discriminative features that are vital for improving expression recognition. The fusion weights of the ASF are generated via the global-local attention, which aggregates global and local context for further expression recognition. The size of fused feature maps X_{fused} is also $H_d \times W_d \times C_f$. Afterwards, we feed the flattened features to a linear projection and a learnable classification token is added. We get embedded visual words with size of $(H_d W_d + 1) \times C_p$, where C_p is the channel of flattened features after projection. We also add position embeddings to the embeddings to retain positional information, as [43] and [49] do. The input embeddings are further fed to the Transformer encoder, which is composed of N_l encoder blocks. Finally, the probabilities of facial expressions are generated by a fully connected layer and the softmax function.

3.2 Attentional Selective Fusion

Our attentional selective fusion consists of global attention and local attention, as can be seen in Fig. 3, which can

provide additional flexibility in fusing different types of information. As mentioned in the Section 3.1, given two feature maps $X_{LBP}, X_{RGB} \in \mathbb{R}^{H_d \times W_d \times C_f}$ extracted from the backbones, we first fuse the LBP features X_{LBP} and the CNN features X_{RGB} for capturing the subsequent information interaction:

$$U = W_L X_{LBP} + W_C X_{RGB}, \tag{1}$$

where U is the integrated feature maps after summation between X_{LBP} and X_{RGB}, and + denotes element-wise summation. W_L and W_C are the weights for initial integration and simply implemented by two 1×1 convolutions. To perform both global and local selective fusion, we then choose global average pooling and the pixel-wise convolution as global context and local context aggregator, respectively. The global context and local context are computed as follows:

$$G(U) = \sigma(\mathbb{BN}(Conv_G^2(\delta(\mathbb{BN}(Conv_G^1(\mathbb{AP}(U))))))), \tag{2}$$

$$L(U) = \sigma(\mathbb{BN}(Conv_L^2(\delta(\mathbb{BN}(Conv_L^1(U)))))), \tag{3}$$

where $G(U) \in \mathbb{R}^{1\times1\times C_f}$ and $L(U) \in \mathbb{R}^{H_d\times W_d\times1}$ represent global fusion and local fusion weights. $\mathbb{AP}$ denotes global adaptive average pooling, which is carried out using Eq. (4).

$$\mathbb{AP}(U) = \frac{1}{H_d W_d} \sum_{i=1}^{H_d} \sum_{i=1}^{W_d} x_c(i, j), c = 1, 2, \ldots, C_f, \tag{4}$$

where H_d and W_d are the height and width of the input feature map U, and C_f is the number of channels of U. $\mathbb{BN}$ is the Batch Normalization. δ denotes the ReLU function, and σ denotes the Sigmoid function. The kernel sizes of $Conv_G^1$ and $Conv_G^2$ are $\frac{C_f}{r} \times C_f \times 1 \times 1$, $C_f \times \frac{C_f}{r} \times 1 \times 1$. We set r to 8 in this paper. Similarly, $L(U) \in \mathbb{R}^{H_d\times W_d\times1}$ is the local fusion weights. And the kernel sizes of $Conv_L^1$, $Conv_L^2$ are $\frac{C_f}{r} \times C_f \times 1 \times 1$, $1 \times \frac{C_f}{r} \times 1 \times 1$. Given the global fusion weights $G(U)$ and the local fusion weights $L(U)$, the refined global-local attention weights can be obtained by Eq. (5).

$$GL(U) = G(U) \oplus L(U), \tag{5}$$

where $\oplus$ represents the broadcasting addition. Then, the fused feature map X_{fused} is calculated by as follows:

$$\begin{aligned} X_{fused} =& X_{LBP} \otimes \sigma(GL(U)) + \\ & X_{RGB} \otimes \sigma(1 - GL(U)), \end{aligned} \tag{6}$$

where $\otimes$ is the element-wise multiplication.

3.3 Multi-Layer Transformer Encoder

The fused 2D feature map X_{fused} need to be flattened into a 1D visual embedding sequence, and further can be fed for the multi-layer Transformer encoder as input. Therefore, we reshape $X_{fused} \in \mathbb{R}^{H_d\times W_d\times C_f}$ into a flattened sequence and feed it to a linear projection to get $X_f \in \mathbb{R}^{H_d W_d\times C_p}$, where $H_d W_d$ is the sequence length and C_p, C_f are set to 512 and 768 respectively. As in [49], a classification token [cls] is appended at the beginning of the input sequence X_f. The learnable state of the [cls] token at the output of the Transformer encoder is utilized to represent the whole feature sequence, which serves for the final prediction. To incorporate the positional information in the multi-layer Transformer encoder, the 1D learnable positional embeddings are added to the feature embeddings:

$$Z^0 = [x_{cls}; x_f^1; x_f^2; x_f^3; \ldots; x_f^{H_d W_d}] + PE(H_d W_d + 1; C_p), \tag{7}$$

where $PE(H_d W_d + 1; C_p) \in \mathbb{R}^{(H_d W_d+1)\times C_p}$ learns the embeddings for each position index, [cls] token included, and Z^0 represents the resulting position-aware feature sequence.

To model the complex interactions among all elements of the facial feature embeddings, we input Z^0 to the standard multi-Layer Transformer encoder. The Transformer encoder calculates the weights of embeddings Z^0 through multi-head self-attention (MHSA). This is done by learnable queries Q, keys K, and values V. We compute the single-head global self-attention(SHSA) using Eq. (8). Details of SHSA in the first layer can be formulated as follows:

$$\begin{aligned} head_j &= Attention(Q_j, K_j, V_j) \\ &= softmax(\frac{Q_j K_j^T}{\sqrt{d}}) V_j \\ &= softmax(\frac{Z^0 W_j^Q (Z^0 W_j^K)^T}{\sqrt{d}}) Z^0 W_j^V, \end{aligned} \tag{8}$$

where $Q_j = Z^0 W_j^Q$, $K_j = Z^0 W_j^K$, $V_j = Z^0 W_j^V$ and $W_j^Q \in \mathbb{R}^{C_p\times d}$, $W_j^K \in \mathbb{R}^{C_p\times d}$, $W_j^V \in \mathbb{R}^{C_p\times d}$ are the parameters of these linear projections. Specifically, multi queries, keys and values project Z^0 into N_h different representation subspaces. Multi-head self-attention (MHSA) can be described as:

$$MHSA(Z^0) = concat(head_1, \ldots, head_{N_h}) W^O, \tag{9}$$

where N_h is the number of different heads, and $concat$ denotes the concatenation operation. $W^O \in \mathbb{R}^{h_1\times d}$ are the parameters of a linear projection, where the dimension of each head d is equal to $\frac{C_p}{N_h}$ and h_1 is the hidden size of the first layer. Each Transformer encoder consists of N_l layers of MHSA blocks. Formally, the standard multi-layer Transformer encoder computes forwardly for $i = 1, \ldots, N_l$ layers:

$$\hat{Z}^i = MHSA(LN(Z^{i-1})) + Z^{i-1} \tag{10}$$

$$Z^i = MLP(LN(\hat{Z}^i)) + \hat{Z}^i \tag{11}$$

where $\hat{Z}^i$ and Z^i are intermediate output and final output at layer i. The MLP consists of two position-wise feed-forward layers and a GELU non-linearity activation function. The LN denotes the layer normalization, which is applied before every attention block and the MLP. The hidden dimension of the MLP is set to 3,072 in this paper. The output $Z_0^{N_l}$ of the layer N is also normalized by the LN. It is notable that we just apply the fully-connected layer for the final [cls] token, which is further for calculating the expression probability scores. Mathematically, the probability scores are generated as follows:

$$Y = LN(Z_0^{N_l}), \tag{12}$$

$$y^i = \frac{e^{\theta_i^T Y}}{\sum_{i=1}^{M} e^{\theta_i^T Y}}, \tag{13}$$

where $Z_0^{N_l}$ is the first [cls] token of the whole sequence Z_l^N and Y is the output of the Multi-layer Transformer encoder. θ represents the parameters of the fully-connected layer and

θ_i is the i-th column of θ. The number of expression classes is M. y^i is the probability score of the i-th facial expression, and the final predicted expressions can be easily obtained by arg max function during inference.

4 Experiments

In order to demonstrate the effectiveness of our proposed method, we carry out extensive experiments on three in-the-wild FER datasets (i.e., RAF-DB, FERPlus and AffectNet) and cross-dataset evaluation on CK+. In this section, we first introduce the FER datasets used in our experiments and implementation details. Then, the proposed method is compared with several state-of-the-art approaches. Subsequently, the impact of each component of the proposed CVT model is investigated with experiments on these datasets.

4.1 Datasets

We evaluate our approach on three frequently used facial expression datasets (RAF-DB, FERPlus and AffectNet). These datasets are all collected in the wild, which may suffer from different illuminations and occlusions. To demonstrate the effectiveness of our method when handling occlusion and variant pose issues in real-word conditions, we also conduct experiments on Occlusion-RAF-DB, Pose-RAF-DB, Occlusion-FERPlus, Pose-FERPlus, Occlusion-AffectNet, Pose-AffectNet, which are subsets of RAF-DB, FERPlus and AffectNet, respectively. The cross-dataset evaluation experiments are also conducted on CK+ to verify the superior generalization ability of our method. The details of the datasets used in the experiments are introduced as follows.

RAF-DB contains 29,672 real-world facial images collected from Flickr. The whole images of RAF-DB are labeled by 315 well-trained annotators and each image is labeled by about 40 independent annotators. RAF-DB contains two different subsets: single-label subset and multi-label subset. In our experiments, we only use single-label subset, including seven basic emotions (neutral, happy, surprise, sad, angry, disgust, fear). The images from the single-label subset are split into 12,271 training samples and 3,068 testing samples. The expressions in both training images and test images have imbalanced distribution. The overall sample accuracy is used for performance measurement.

FERPlus is extended from the original FER2013 dataset, which was for the ICML 2013 Challenges in Representation Learning. FERPlus contains 28,709 training images and 3,589 test images, all of which are collected by the Google search engine. The original size of the images in FERPlus is 48 $\times$ 48. Each image in FERPlus is annotated by 10 annotators and FERPlus provides better quality labels than the original FER2013 labels. Apart from seven basic emotions as RAF-DB, the contempt category is included in the labels. We mainly report overall sample accuracy under the supervision of majority voting for performance measurement.

AffectNet is the largest facial expression datasets with more than 1,000,000 facial images collected from the Internet. AffectNet provides both discrete categorical and continuous dimensional (i.e., valence and arousal) annotations. It should be noted that AffectNet has imbalanced training, validation and test sets, of which 450,000 images have been annotated manually. In our experiment, we utilize images annotated with eight basic expressions as FERPlus, 287,652 images for training and 4,000 images for testing. Since the test set is not available to the public, we mainly report mean class accuracy on the validation set for performance measurement and fair comparison with other methods.

Occlusion and Pose Variant Datasets (i.e., Occlusion-RAF-DB, Pose-RAF-DB, Occlusion-FERPlus, Pose-FERPlus, Occlusion-AffectNet, Pose-AffectNet) are occlusion and pose subsets of RAF-DB, FERPlus, AffectNet and manually collected by [10]. There are various occlusion types occurring in samples of these datasets, such as wearing mask, wearing glasses and objects in upper/bottom face. In addition, variant pose issues can be divided into two categories, poses larger than 30 degrees and poses larger than 45 degrees. The sample distribution can be found in [10]. We report overall sample accuracy or mean class accuracy according to corresponding original datasets.

CK+ is the extended Cohn-Kandade(CK) dataset for facial action unit and expression recognition, collected under a lab-collected environment. The original data of CK+ has 593 video sequences from 123 subjects, 327 of which are annotated with seven basic emotions and contempt. Each of the video sequences consists of images from onset (the first frame) to peak expression (the last frame). We follow previous work to utilize the first frame of a video sequence as a neutral sample and the last frame with the target expression. In total, we obtain 618 images with seven emotions and 654 images with eight emotions for testing. We use the overall sample accuracy to evaluate the generalization capacity of our method.

4.2 Implementation Details

In our experiments, the face images are detected by MTCNN [19] and further resized to the size of 224 $\times$ 224. For fair comparison with previous state-of-the-art methods, we use the same backbone ResNet18 pre-trained on the MS-Celeb-1M face recognition dataset. The facial features are extracted from the last convolutional stage of ResNet18. The learning rate of our methods is initialized as 0.005. We use a linear learning rate warmup of 1,000 steps and cosine learning rate decay. The Adam optimizer [58] is used to optimize the whole networks with a batch size of 32 and train the model for 20,000 steps on RAF and FERPlus, 40,000 steps on AffectNet, respectively. The standard cross-entropy loss is utilized to supervise the model to generalize well for expression recognition. We implement our method with Pytorch [59] toolbox and conduct all the experiments on a single NVIDIA GTX 1080Ti GPU card.

4.3 Comparison with State-of-the-art Methods

The proposed CVT model is compared with several state-of-the-art methods on RAF-DB, FERPlus and AffectNet. We achieve new state-of-the-art results on these datasets to our knowledge. The better performance demonstrates the superiority of our proposed method.

Results on RAF-DB: Comparison with other state-of-the-art methods can be found in Table 1. The methods in [8] presented their performance using mean accuracy. For fair

TABLE 1
Comparison with state-of-the-art methods on RAF-DB. The best results are in bold.

method	Year	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Accuracy
VGG [8]	2018	66.05	25.00	37.84	73.08	51.46	53.49	47.21	69.34
baseDCNN [8]	2018	70.99	52.50	50.00	92.91	77.82	79.64	83.09	82.66
Center Loss [8]	2018	68.52	53.13	54.05	93.08	78.45	79.63	83.24	82.86
DLP-CNN [8]	2018	71.60	52.15	62.16	92.83	80.13	81.16	80.29	82.74
FSN [52]	2018	72.80	46.90	56.80	90.50	81.60	81.80	76.90	81.14
gACNN [9]	2018	-	-	-	-	-	-	-	85.07
RAN [10]	2020	-	-	-	-	-	-	-	86.90
SCN [11]	2020	-	-	-	-	-	-	-	87.03
DSAN-VGG-RACE [24]	2020	82.71	56.25	58.11	94.01	83.89	**89.06**	80.00	85.37
SPWFA-SE [53]	2020	80.00	59.00	59.00	93.00	84.00	88.00	86.00	86.31
Ours	2021	**85.80**	**68.12**	**64.86**	**94.09**	**87.24**	85.41	**87.50**	**88.14**

TABLE 2
Comparison with state-of-the-art methods on FERPlus and AffectNet. The best results are in bold.

(a) Results on FERPlus.

Method	Year	Accuracy
CSLD [15]	2016	83.85
ResNet+VGG [54]	2017	87.4
SHCNN [55]	2019	86.54
LDR [56]	2020	87.6
RAN$^\diamond$ [10]	2020	88.55
RAN [10]	2020	87.85
SCN [11]	2020	88.01
Ours	2021	**88.81**

(b) Results on AffectNet.

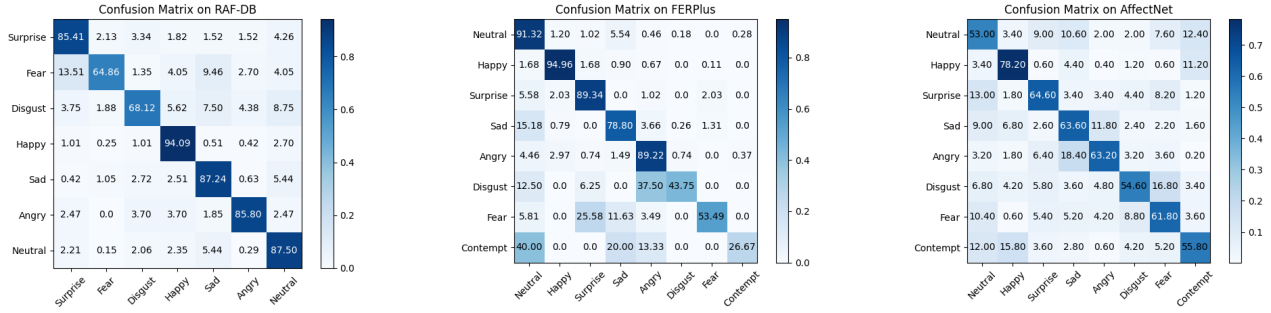
Method	Year	Accuracy
IPA2LT [57]	2018	55.11
gACNN [9]	2018	58.78
SPWFA-SE [53]	2020	59.23
RAN* [10]	2020	52.97
RAN [10]	2020	59.50
SCN [11]	2020	60.23
Ours	2021	**61.85**

comparison, we convert their results to accuracy, following [53]. Since gACNN [9], RAN [10] and SCN [11] did not report the specific expression recognition accuracy or the confusion matrices, the corresponding places are marked with '-' in Table 1. Although SPWFA-SE [53] also did not report specific expression recognition accuracy, it provided the confusion matrix on RAF-DB. Therefore, we borrow the accuracy results from its confusion matrix for comparison. Overall, our proposed method achieves 88.14% on RAF-DB, and the corresponding confusion matrix is shown in Fig. 4(a). As shown in Table 1, our method achieves all the best results among all methods except for the surprise category. In detail, our CVT has obtained gains of 18.8% and 1.11% over VGG and SCN, which are the baseline method and the previous state-of-the-art method, respectively. DSAN-VGG-RACE integrated deeply-supervised blocks and attention blocks with race labels, which were additional data compared with our purely used expression labels. Since RAF-DB also has extremely imbalanced distribution, the slight performance decline of the surprise category is reasonable and acceptable. The accuracy of the disgust expression recognition of our method has recorded an increase of 9.12% compared with the previous best result of [53], which demonstrates the effectiveness and the superiority of our method in feature learning.

Results on FERPlus: Table 2(a) presents the comparison results on FERPlus. We compare our models with CNN-based methods including CSLD [15], ResNet+VGG [54], SHCNN [55], LDR [56], as well as two recent state-of-the-art methods(RAN [10], SCN [11]). As we can see in Table 2(a), our proposed CVT achieves 88.81% on FERPlus. Under the same experiment settings, total improvements of our CVT on FERPlus are 0.96% and 0.80% when compared with RAN and SCN. Especially, RAN$^\diamond$ means the improved RAN with extra face alignment. Although face alignment is crucial for face recognition and facial expression recognition, it is a preprocessing step and CNN-based methods tend to recognize expression in an end-to-end manner. Even without face alignment in our experiment settings, our CVT also achieves much better results over RAN$^\diamond$.

Results on AffectNet: We compare our method with several methods on AffectNet and report the results in Table 2(b). We obtain 61.85% with oversampling on AffectNet, without bells and whistles. IPA2LT [57], gACNN [9] and SPWFA-SE [53] are trained for seven classes on AffectNet without the contempt category. As mentioned above, AffectNet has imbalanced distribution. SPWFA-SE utilizes focal loss function to handle the imbalance problem. To deal with the imbalance issue, as RAN [10] and SCN [11] do, we adopt the oversampling strategy in our experiments. Especially, RAN$^\diamond$ denotes RAN without using the oversampling strategy for training. Our CVT with only AffectNet for training outperforms SCN by 1.62%, which applied extra dataset WebEmotion for pre-training and then fine-tuned SCN on AffectNet. The improvements of our CVT over previous methods suggest that the CVT indeed has better generalization ability even on large-scale expression recognition datasets like AffectNet.

Results on Occlusion and Pose Variant Datasets: To examine our method in case of occlusion and variant pose in real scenarios, we also conduct several experiments on Occlusion-RAF-DB, Pose-RAF-DB, Occlusion-FERPlus, Pose-FERPlus, Occlusion-AffectNet and Pose-AffectNet. Table 3 shows the accuracy of the experimental results under corresponding subsets. RAN [10] proposed to divide a face image into subregions and introduced a region biased loss



(a) Confusion Matrix of our CVT on RAF-DB. (b) Confusion Matrix of our CVT on FERPlus. (c) Confusion Matrix of our CVT on AffectNet.

Fig. 4. The confusion matrices of our method on RAF-DB, FERPlus and AffectNet. The diagonal values of each confusion matrix corresponds the accuracy of specific expressions. The darker the cell, the higher its accuracy.

TABLE 3
Comparison with other methods on Occlusion and Pose Variant Datasets.

(a) Results on Occlusion-RAF-DB, Pose-RAF-DB.

Method	Occlusion	Pose(30)	Pose(45)
Baseline [10]	80.19	84.04	83.15
RAN [10]	82.72	86.74	85.20
Ours	83.95	87.97	88.35

(b) Results on Occlusion-FERPlus, Pose-FERPlus.

Method	Occlusion	Pose(30)	Pose(45)
Baseline [10]	73.33	78.11	75.50
RAN [10]	83.63	82.23	80.40
Ours	84.79	88.29	87.20

(c) Results on Occlusion-AffectNet, Pose-AffectNet.

Method	Occlusion	Pose(30)	Pose(45)
Baseline [10]	49.48	50.10	48.50
RAN [10]	58.50	53.90	53.19
Ours	62.98	60.61	61.00

TABLE 4
Cross-dataset evaluation results on CK+.

Method	Train	Test	Accuracy
gACNN [9]	RAF-DB	CK+	81.07
SPWFA-SE [53]	RAF-DB	CK+	81.72
SPWFA-SE [53]	AffectNet	CK+	85.44
Ours	RAF-DB	CK+	81.88
Ours	FERPlus	CK+	83.79
Ours	AffectNet	CK+	86.24

for capturing the importance of different regions for occlusion and pose variant expression recognition. Although our CVT is not specifically designed for occlusion and variant pose FER issues, our method outperforms RAN with a large margin in each case, which shows the superiority of our method. Specifically, our method exceeds RAN by 1.23%, 1.16% and 4.48% on Occlusion-RAF-DB, Occlusion-FERPlus and Occlusion-AffectNet. Our method also outperforms RAN on Pose-RAF-DB, Pose-FERPlus and Pose-AffectNet. The gains are 1.23%, 6.06% and 6.71% with pose larger than 30 degrees. On Pose-FERPlus, Pose-AffectNet and Pose-RAF-DB with pose larger than 45 degrees, our method significantly outperforms RAN with the gains of 3.15%, 6.8% and 7.81%, respectively. Overall, these results reliably verify the effectiveness of our method on occlusion and variant pose issues. In addition, the superior performance consists with our hypothesis that it is feasible and effective to recognize the facial expressions by a sequence of visual words from a global perspective.

Results on CK+: We also conduct cross-dataset evaluation to verify the superior generalization ability of our method. Specifically, we first train the network individually on RAF-DB, FERPlus as well as AffectNet, and then evaluate the model directly on CK+. Table 4 shows that our method achieves better performance than previous approaches. Note that gACNN and SPWFA-SE are trained for predicting seven basic emotions. Although our model predicts one more expression (contempt), which generally lowers down the final results as shown in Fig. 4(b) and Fig. 4(c), the model trained on AffectNet gets an accuracy of 86.54% on CK+. Compared with the gACNN and the SPWFA-SE, our method has better performance and increases by 0.81% and 0.16%. The model trained on AffectNet achieves higher accuracy than the ones trained on RAF-DB and FERPlus, because AffectNet is a relatively large scale dataset and contains more facial images. Table 4 demonstrates that our method has better generalization capacity and achieves better performance without exceptions.

4.4 Ablation Study

As shown in Fig 2, our proposed method CVT consists of LBP features with the attentional selective fusion (ASF) module and the multi-layer Transformer encoder (MTE). To validate the effectiveness of these modules, we conduct comparative experiments on RAF, FERPlus and AffectNet by discarding some parts of our CVT. The detail settings of these experiments can are found in Table 5, where the setting (**a**) represents the baseline method. Since the ASF is used to integrate the LBP features and the CNN features, it can not be retained without the LBP module.

Effectiveness of the LBP features for FER. Since our method begins with extracting the LBP features, we design

TABLE 5
Ablation study w.r.t. the LBP features, the ASF and the MTE, performed on RAF-DB, FERPlus and AffectNet.

Setting	LBP	ASF	MTE	RAF-DB	FERPlus	AffectNet
a	✗	✗	✗	86.43	86.75	58.41
b	✓	✗	✗	86.62	87.10	58.67
c	✓	✓	✗	87.23	87.48	59.80
d	✗	✗	✓	87.57	87.63	60.48
e	✓	✗	✓	87.84	88.11	60.95
f	✓	✓	✓	88.14	88.69	61.55

TABLE 6
Ablation study w.r.t. the number of heads, the number of layers, performed on RAF-DB, FERPlus and AffectNet. Bold values correspond to the best performance.

Setting	N_l	N_h	Params(M)	RAF-DB	FERPlus	AffectNet
i	4	4	51.8	87.45	88.46	61.08
ii	4	8	51.8	**88.14**	88.69	61.55
iii	4	12	51.8	87.52	88.65	**61.85**
iv	8	4	80.1	87.22	88.14	60.82
v	8	8	80.1	87.61	**88.81**	61.23
vi	8	12	80.1	87.48	88.52	61.30
vii	12	4	108.5	87.23	88.20	60.85
viii	12	8	108.5	87.29	88.52	60.45
ix	12	12	108.5	87.09	88.21	61.50

the ablation study to investigate the impact of LBP features for FER. Taking results on RAF-DB for example, the baseline on RAF-DB is 86.43 % without any modifications. Adding the LBP features on RAF-DB improves the baseline by 0.19%. Settings (**a**, **b**) and settings (**d**, **e**) demonstrate that integrating the LBP features improves the baselines on FERPlus and AffectNet by 0.25% and 0.26%, which also suggests the LBP features are beneficial in improving expression recognition performance. This can be explained by that the LBP features can extract texture information and reflect fine facial changes, which show the subtle differences of expressions. Nevertheless, directly using additional the LBP features for FER is of limited use, because the simple addition fusion strategy is unsatisfactory for combining the LBP features and the CNN features.

Evaluation of the attentional selective fusion (ASF). To verify the effectiveness of the ASF for fusing LBP features and CNN features, we conduct experiments by replacing the ASF with the simple element-wise addition. According to settings (**b**, **c**) and settings (**e**, **f**), the ASF leads to an increase in recognition accuracy when fusing the LBP features and the CNN features, showing the effectiveness of the proposed ASF. Specifically, we can see from settings (**b**, **c**) that the designed ASF further improves the performance by 0.61%, 0.38% and 1.27%. According to the settings (**a**, **c**) the ASF improves the baseline setting (**a**) by 0.8%, 0.73% and 1.39% with additional LBP features, respectively. The ASF aggregates global and local contexts for fusing the LBP features and the CNN features, which further improves the recognition performance.

Effectiveness of the multi-layer Transformer encoder (MTE). To explore the impact of the MTE, we evaluate the performance of the MTE in this part. We also compare the effects of the MTE on RAF-DB, FERPlus and AffectNet. For fair comparison, we constrain the number of heads N_h and the number of layers N_l to 4 and 8, respectively. From settings (**a**, **d**), settings (**b**, **e**) and settings (**c**, **f**), we can clearly see that the MTE greatly improves the performance. Specifically, compared with the baseline setting (**a**), the setting (**d**) shows that integrating the MTE outperforms the baseline by 1.24%, 0.88% and 2.07% on RAF-DB, FERPlus and AffectNet. The MTE contributes most to the accuracy improvements over the LBP features and the ASF. We infer that employing the MTE increases the ability of learning discriminative features, outperforming corresponding baselines.

TABLE 7
Comparison between ViT and our CVT.

Setting	N_l	N_h	Params(M)	RAF-DB	FERPlus	AffectNet
ViT	12	12	85.8	47.55	47.72	27.87
ViT*	12	12	85.8	85.14	88.07	58.77
Ours	4	8	51.8	82.27	84.80	58.75
Ours*	4	8	51.8	88.14	88.69	61.55

Impact of the number of layers and heads of the MTE. The multi-layer Transformer encoder consists of N_l identical layers. The multi-head self-attention in each layer enables the model decompose the information into N_h representation subspaces and jointly capture discriminative information at different positions. We explore the effects of different layer values N_l and the number of heads N_h from 4 to 12 on RAF-DB, FERPlus and AffectNet. Table 6 compares the performance of different hyper-parameter settings of our method in terms of accuracy/mean accuracy and parameters. We observe that increasing the number of layers N_l greatly burdens the parameters of the whole network. Generally, smaller N_l values tend to achieve better recognition performance, and larger N_l values result in excessive parameters at the risk of overfitting. Note that the smaller the value of N_h, the worse performance we may get, because there are not enough subspaces to learn latent representations.

Impact of the pre-trained weights. To find out the impact of pre-trained weights on our method, we train our model from scratch or fine-tune the model from pre-trained ResNet18 weights. We also give a comparison of ViT and our CVT in Table 7. We implement ViT-Base [49] on these three FER datasets and conduct experiments based on default settings as [49] described. The ViT denotes that it is trained from scratch and ViT* represents we fine-tune it with weights pre-trained on ImageNet-21k and ImageNet. From settings ViT* and Ours*, we can infer that our method achieves better performance but with fewer parameters. The performance of ViT greatly drops when trained from scratch instead of fine-tuning, because the feature extraction capacity of ViT is relatively limited without the guidance of large-scale datasets. Table 7 shows that fine-tuning models from pre-trained weights usually results in better performance.

4.5 Visualization

Our method computes relationships between visual words and captures discriminative features for expression recog-

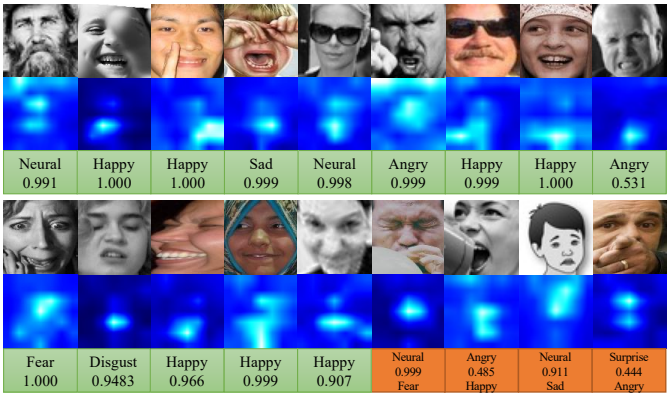


Fig. 5. An illustration of the learned attention maps. The first row is to show the raw images, the second row is the attention maps of the MTE, and the corresponding predictions or the labels are shown in the third row. The annotations marked in green are the predicted expressions and their confidence scores, while the others in orange are false predictions with the ground-truth labels annotated.

nition. Table 3 provides an empirical evidence to support the effectiveness of our method. To visually demonstrate the superior capacity of our method, the raw images and their corresponding attention weight maps are visualized in Fig. 5. We randomly select several images from RAF-DB, FERPlus and AffectNet, and compute the attention weight maps across all heads. The weights of all layers are then multiplied recursively and projected into the input image space.

As shown in Fig. 5, the weight maps shows that our method can identify occlusion regions and highlight discriminative features. For example, the third raw image is occluded with a finger on the left part, while our method provides high attention weights on the right part. The weight map of the eighth raw image contains high attention in the bottom, covering the mouth area. We also provide some false predictions in Fig. 5. The first false-positive sample indicates that a man is blowing up a balloon about to burst and his eyes close out of fear. It shows that the detection process is a double-edged sword, which provides the precious face locations but excludes the background knowledge. The cartoon face is unfortunately predicted incorrectly because of limited cartoon faces in these datasets. The other two false-positive samples also indicate the external knowledge like the social norms should be taken into consideration for the future research work. From Fig. 5, We can conclude that our method can dynamically model the relationships among these visual words and highlight discriminative regions to boost recognition performance.

5 Conclusion

In this paper, we present Convolutional Visual Transformers (CVT) for facial expression recognition in the wild. We propose to tackle the expression recognition problem by translating facial images into sequences of visual words and performing recognition from a global perspective. To achieve these goals, we design the attentional selective fusion to dynamically and adaptively combine LBP features and CNN features for improving recognition accuracy. The visual words are generated by flattening and projecting the fused feature maps. The multi-layer Transformer encoder utilizes the global self-attention mechanism to shift attention to discriminative visual words from a global perspective. The experimental results demonstrate that the CVT exceeds other state-of-the-art methods on three frequently used facial expression datasets, i.e., RAF-DB, FERPlus, AffectNet. The cross-dataset evaluation on CK+ also demonstrates the promising generalization ability of our method.

References

[1] C. Darwin and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998. 1

[2] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97–115, 2001. 1

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893. 1

[4] C. Shan, S. Gong, and P. W. McOwan, "Robust facial expression recognition using local binary patterns," in *IEEE International Conference on Image Processing*, vol. 2, 2005, pp. II–370. 1

[5] X. Feng, M. Pietikainen, and A. Hadid, "Facial expression recognition with local binary patterns and linear programming," *Pattern Recognition And Image Analysis C/C of Raspoznavaniye Obrazov I Analiz Izobrazhenii*, vol. 15, no. 2, p. 546, 2005. 1

[6] I. Buciu and I. Pitas, "Application of non-negative and local non negative matrix factorization to facial expression recognition," in *Proceedings of the 17th International Conference on Pattern Recognition*, vol. 1, 2004, pp. 288–291. 1

[7] S. H. Lee, K. N. Plataniotis, and Y. M. Ro, "Intra-class variation reduction using training expression images for sparse representation based facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 340–351, 2014. 1

[8] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2018. 1, 2.1, 4.3, 1

[9] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2018. 1, 2.1, 1, 2(b), 4.3, 4.3, 4

[10] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020. 1, 2.1, 4.1, 1, 2(a), 2(b), 4.3, 4.3, 4.3, 3(a), 3(b), 3(c)

[11] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6897–6906. 1, 2.1, 1, 2(a), 2(b), 4.3, 4.3

[12] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *IEEE Conference on Computer Vision and Pattern Recognition-workshops*, 2010, pp. 94–101. 1

[13] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," in *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, 2010, pp. 65–70. 1

[14] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011. 1

[15] E. Barsoum, C. Zhang, C. C. Ferrer, and 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*, 2016, pp. 279–283. 1, 2(a), 4.3

[16] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017. 1

[17] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989. 1

[18] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, 2020, doi: 10.1109/TAFFC.2020.2981446. 1

[19] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016. 1, 4.2

[20] B. Amos, B. Ludwiczuk, M. Satyanarayanan *et al.*, "Openface: A general-purpose face recognition library with mobile applications," *CMU School of Computer Science*, vol. 6, no. 2, 2016. 1

[21] F. Bourel, C. C. Chibelushi, and A. A. Low, "Recognition of facial expressions in the presence of occlusion." in *Procedings of the British Machine Vision Conference*, 2001, pp. 1–10. 1

[22] S. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 1–12, 2014. 1

[23] Y. Li, J. Zeng, S. Shan, and X. Chen, "Patch-gated cnn for occlusion-aware facial expression recognition," in *24th International Conference on Pattern Recognition*, 2018, pp. 2209–2214. 1

[24] Y. Fan, V. Li, and J. C. Lam, "Facial expression recognition with deeply-supervised attention network," *IEEE Transactions on Affective Computing*, 2020, doi: 10.1109/TAFFC.2020.2988264. 1, 2.1, 1

[25] J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju, "Attention mechanism-based cnn for facial expression recognition," *Neurocomputing*, vol. 411, pp. 340–350, 2020. 1, 2.2

[26] W. Zheng, "Multi-view facial expression recognition based on group sparse reduced-rank regression," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 71–85, 2014. 1

[27] Y. Liu, J. Zeng, S. Shan, and Z. Zheng, "Multi-channel pose-aware convolution neural networks for multi-view facial expression recognition," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 458–465. 1

[28] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018. 1

[29] Y.-H. Lai and S.-H. Lai, "Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 263–270. 1

[30] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Geometry guided pose-invariant facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4445–4460, 2020. 1

[31] N. Sun, Q. Lu, W. Zheng, J. Liu, and G. Han, "Unsupervised cross-view facial expression image generation and recognition," *IEEE Transactions on Affective Computing*, 2020. 1

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. 2.1, 3.1

[33] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708. 2.1

[34] Y. Tang, "Deep learning using linear support vector machines," *arXiv preprint arXiv:1306.0239*, 2013. 2.1

[35] T. Xu, J. White, S. Kalkan, and H. Gunes, "Investigating bias and fairness in facial expression recognition," in *European Conference on Computer Vision*, 2020, pp. 506–523. 2.1

[36] K.-H. Pong and K.-M. Lam, "Multi-resolution feature fusion for face recognition," *Pattern Recognition*, vol. 47, no. 2, pp. 556–567, 2014. 2.2

[37] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125. 2.2

[38] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1925–1934. 2.2

[39] Y. Hu, Z. Zeng, L. Yin, X. Wei, X. Zhou, and T. S. Huang, "Multi-view facial expression recognition," in *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, 2008, pp. 1–6. 2.2

[40] S. Moore and R. Bowden, "Local binary patterns for multi-view facial expression recognition," *Computer Vision and Image Understanding*, vol. 115, no. 4, pp. 541–558, 2011. 2.2

[41] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Facial expression recognition in video with multiple feature fusion," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 38–50, 2016. 2.2

[42] J. Shao and Y. Qian, "Three convolutional neural network models for facial expression recognition in the wild," *Neurocomputing*, vol. 355, pp. 82–92, 2019. 2.2

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017. 2.3, 3.1

[44] J. Beal, E. Kim, E. Tzeng, D. H. Park, A. Zhai, and D. Kislyuk, "Toward transformer-based object detection," *arXiv preprint arXiv:2012.09958*, 2020. 2.3

[45] S. Yang, Z. Quan, M. Nie, and W. Yang, "Transpose: Towards explainable human pose estimation by transformer," *arXiv preprint arXiv:2012.14214*, 2020. 2.3

[46] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," *arXiv preprint arXiv:2012.09841*, 2020. 2.3

[47] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, "End-to-end video instance segmentation with transformers," *arXiv preprint arXiv:2011.14503*, 2020. 2.3

[48] M. Bhat, J. Francis, and J. Oh, "Trajformer: Trajectory prediction with local self-attentive contexts for autonomous driving," *arXiv preprint arXiv:2011.14910*, 2020. 2.3

[49] 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 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. 2.3, 3.1, 3.3, 4.4

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. 2.3

[51] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," *arXiv preprint arXiv:2102.12122*, 2021. 2.3

[52] S. Zhao, H. Cai, H. Liu, J. Zhang, and S. Chen, "Feature selection mechanism in cnns for facial expression recognition." in *Procedings of the British Machine Vision Conference*, 2018, p. 317. 1

[53] Y. Li, G. Lu, J. Li, Z. Zhang, and D. Zhang, "Facial expression recognition in the wild using multi-level features and attention mechanisms," *IEEE Transactions on Affective Computing*, 2020. 1, 2(b), 4.3, 4.3, 4

[54] C. Huang, "Combining convolutional neural networks for emotion recognition," in *IEEE MIT Undergraduate Research Technology Conference (URTC)*, 2017, pp. 1–4. 2(a), 4.3

[55] S. Miao, H. Xu, Z. Han, and Y. Zhu, "Recognizing facial expressions using a shallow convolutional neural network," *IEEE Access*, vol. 7, pp. 78 000–78 011, 2019. 2(a), 4.3

[56] X. Fan, Z. Deng, K. Wang, X. Peng, and Y. Qiao, "Learning discriminative representation for facial expression recognition from uncertainties," in *2020 IEEE International Conference on Image Processing*, 2020, pp. 903–907. 2(a), 4.3

[57] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 222–237. 2(b), 4.3

[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 4.2

[59] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019. 4.2