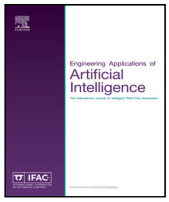
Engineering Applications of Artificial Intelligence 133 (2024) 108535

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



Research paper

Dynamic facial expression recognition based on spatial key-points optimized region feature fusion and temporal self-attention

Zhiwei Huang [a], Yu Zhu [a,*], Hangyu Li [a], Dawei Yang [b,c,*]

[a] School of Information Science and Engineering, East China University of Science and Technology, Shanghai, 200237, PR China
[b] Department of Pulmonary and Critical Care Medicine, Zhongshan Hospital, Fudan University, Shanghai, 200032, PR China
[c] Shanghai Engineering Research Center of Internet of Things for Respiratory Medicine, Shanghai, 200032, PR China

ARTICLE INFO

Keywords:
Dynamic facial expression recognition
Spatial feature fusion
Graph convolution network
Self-attention

ABSTRACT

Dynamic facial expression recognition (DFER) is of great significance in promoting empathetic machines and metaverse technology. However, dynamic facial expression recognition (DFER) in the wild remains a challenging task, often constrained by complex lighting changes, frequent key-points occlusion, uncertain emotional peaks and severe imbalanced dataset categories. To tackle these problems, this paper presents a depth neural network model based on spatial key-points optimized region feature fusion and temporal self-attention. The method includes three parts: spatial feature extraction module, temporal feature extraction module and region feature fusion module. The intra-frame spatial feature extraction module is composed of the key-points graph convolution network (GCN) and a convolution network (CNN) branch to obtain the global and local feature vectors. The newly proposed region fusion strategy based on face spatial structure is used to obtain the spatial fusion feature of each frame. The inter-frame temporal feature extraction module uses multi-head self-attention model to obtain the temporal information of inter-frames. The experimental results show that our method achieves accuracy of 68.73%, 55.00%, 47.80%, and 47.44% on the DFEW, AFEW, FERV39k, and MAFW datasets. Ablation experiments showed that the GCN module, fusion module, and temporal module improved the accuracy on DFEW by 0.68%, 1.66%, and 3.25%, respectively. The method also achieves competitive results in terms of parameter quantity and inference speed, which demonstrates the effectiveness of the proposed method.

1. Introduction

Great advances have been achieved in automated facial expression recognition (FER) based on deep learning (Zhao and Pietikainen, 2007; Dhall et al., 2013; Wang et al., 2020a). Facial expression recognition has great application prospects in human–computer interaction, intelligent assisted driving, psychological medicine and business fields. Till now, a large number of facial expression recognition methods have been proposed (Kossaifi et al., 2020; Zheng et al., 2023; Liu et al., 2023). Dynamic Facial Expression Recognition (DFER) aims to distinguish the emotional categories of the target subject from a continuous video sequence. Compared to Static Facial Expression Recognition (SFER), DFER in the wild needs to address three key issues: inconsistent expressions in one sample, blurry peak frame and facial defects in samples.

Expression inconsistency refers to the presence of emotions in certain frames of the sample that do not match the overall label. Fig. 1(a) shows a sample from a in-the-wild DFER dataset. The label on the left representing the overall expression label and labels below are inferred by the a SFER model, which means that the model needs to focus on the entire sequence rather than a single frame to overcome the issue of inconsistent sentiment categories in the video sequence. Sample peak frame blurring refers to the lack of emotional peak frame annotations in the dynamic expression dataset of natural scenes, and the scattered peak frames in the samples, which increases the difficulty compared to the controlled laboratory dataset (Lucey et al., 2010; Taini et al., 2008; Pantic et al., 2005). As shown in Fig. 1(b), the model needs to overcome the impact of these defects on discrimination accuracy through global and local features.

In addition, video sequences can provide more information with intra-frame and inter-frame. Some static information between inter-frames is redundant for the high sampling frame rate, so most of the DFER methods (Wang et al., 2020a; Meng et al., 2019a) first select a certain number of frames as sequential input frames. Then extract the spatial expression features of each frame separately with convolutional

* Corresponding authors.
E-mail addresses: zhuyu@ecust.edu.cn (Y. Zhu), yang_dw@hotmail.com (D. Yang).

https://doi.org/10.1016/j.engappai.2024.108535
Received 7 November 2022; Received in revised form 1 April 2024; Accepted 29 April 2024
Available online 16 May 2024
0952-1976/© 2024 Elsevier Ltd. All rights reserved.

Fig. 1. Dynamic facial expression recognition dataset samples **(a)** Expression inconsistency and blurry peak frame issues, the red box is the emotion peak frames **(b)** Facial defects in the samples.

neural networks (CNN) (He et al., 2016) or ViT (Dosovitskiy et al., 2020), and finally mine the temporal expression information between the features of sequential frames.

In the field of computer vision, the classical network for exploring temporal information is the recurrent convolutional network (RNN) (Zaremba et al., 2014), which can explore temporal information from a sequence and output the sequence features. However, RNN cannot be parallelized. Therefore, self-attentive models (Vaswani et al., 2017) that are more effective and can be parallelized have gained the attention of researchers. The earliest self-attentive models were proposed in the field of natural language processing and were widely used. In the last two years, the self-attentive model ViT has been introduced into computer vision with good results. Inspired by this, we adopt the self-attentive model to explore the temporal expression information for videos.

The DFER framework proposed in this paper first select a certain number of frames from a video as input frame sequence. Like static face images, face videos in the wild is more likely to have interference factors such as occlusion and side faces. To reduce the influence of these factors, we use a method based on graph convolution network for intra-frame feature extraction, and then propose an effective spatial region feature fusion method to aggregate the facial region and global features as tokens which are fed into temporal self-attention module to represent the inter-frame information for DFER.

A new framework for DFER in-the-wild is proposed. Our major contributions are listed as follows:

- The spatial feature extraction module utilizes facial key-points and graph convolutional network (GCN) to enhance the spatial features, which can overcome occlusion and defects in videos;
- We propose a novel region feature fusion module which can achieve feature aggregation by considering the physical meaning of facial key-points
- The inter-frame temporal module utilizes multi-head self attention mechanism to extract and enhance spatial–temporal features, obtain a discriminative feature with better performance, and improve the classification accuracy;
- We conduct evaluation on four DFER datasets and our method achieves recognition WAR of 68.73%, 55.00%, 47.80%, and 47.44% on the DFEW (Jiang et al., 2020), AFEW (Dhall et al., 2018), FERV39k (Wang et al., 2022a), and MAFW (Liu et al., 2022a) datasets respectively. In addition, it performs well in parameter quantity and inference speed, proving the competitiveness in DFER tasks.

The paper is organized as follows. Section 2 describes the related work in dynamic facial expression recognition, self-attention and graph convolutional network. Section 3 introduces the proposed spatial feature extraction network, fusion module and temporal feature extraction network in detail. In Section 4, we report the experimental datasets and experimental results. The conclusion of this paper is given in Section 5.

2. Related work

2.1. Dynamic facial expression recognition (DFER)

Traditional methods for face expression recognition in dynamic videos are mainly based on manual productions, such as local binary patterns (LBP) based on three orthogonal planes (Zhao and Pietikainen, 2007), vector gradient histograms based on three orthogonal planes (Chen et al., 2014) and spatio-temporal local single gene binary patterns (Huang et al., 2014).

Early deep learning methods mainly include frame based method and temporal based method (Bargal et al., 2016; Kahou et al., 2013). One frame-based method is to aggregate the network output of video sequence frames through various methods, that is, frame aggregation. The other frame based methods (Zhao et al., 2016; Kim et al., 2017)are to design the network according to the peak frame. DenseNet (Liu et al., 2018) is a densely connected network that typically serves as a feature extraction backbone network for visual tasks. Temporal series-based methods mostly use cyclic convolution network and 3D convolution network. RNN and LSTM for DFER (Ebrahimi Kahou et al., 2015; Baddar and Ro, 2019; Lee et al., 2019) first obtained the features of each frame through the basic convolutional neural network (CNN), and then used the cyclic convolution network and LSTM structure to explore the temporal relationship between these features and obtain more expression information.

Newly proposed methods use deeper networks and more unique entry points to improve the accuracy of DFER. EC-STFL (Jiang et al., 2020) addresses the problem of feature edge blurring and sample imbalance through an Expression-Clustered Spatiotemporal Feature Learning framework and a new EC-STFL loss. Kossaifi et al. (2020) proposed a tensor decomposition framework for higher-order multidimensional (separable) convolution, which compresses the network to reduce the number of parameters to improve efficiency, thus alleviating the computational burden one has to bear using 3D spatial–temporal convolutional networks or higher-order multidimensional convolution. CE-FLNet (Liu et al., 2022b) network attempts to find emotional peak segments by segmenting the input video samples, and then optimizes the overall classification results of the samples based on the classification results of emotional peak segments. NR-DFERNet (Li et al., 2022) explored the effectiveness of using inter frame differences in DFER tasks using unique frame difference features and category suppression losses, achieving good results. DPCNet (Wang et al., 2022b) designed a dual stream recognition network that achieved better results on both laboratory datasets and in-the-wild datasets. EST (Liu et al., 2023) model not only uses the Transformer network to fuse the temporal relationships between fragments, but also designs a prediction task to restore unordered fragments to improve the model's temporal prediction ability. ESTLNet (Gong et al., 2024) improves recognition accuracy

Table 1
Previous studies on in-the-wild DFER task.

Method	Publish year	Inputs	Contribution
EC-STFL (Jiang et al., 2020)	2020	TI	Propose with DFEW dataset and raise an effective module for backbone networks in DFER.
Former-DFER (Zhao and Liu, 2021)	2021	DS	A novel three-part architecture of spacial, temporal and classify with Transformer blocks.
STT (Ma et al., 2022)	2022	DS	Combine spatial–temporal attention in one block
CEFLNet (Liu et al., 2022b)	2022	Clip	Propose a strategy to focus on peak segments and eliminate the influence of irrelevant segments.
NR-DFER (Li et al., 2022)	2022	DS	Using inter-frame feature differences to suppress sub-high category and enhance the highest expression.
DPCNet (Wang et al., 2022b)	2022	DS	Adopting a dual structure and a new dual loss forces model to make consistent predictions for both branches.
T-ESFL (Liu et al., 2022a)	2022	Multimodal	Using multimodal inputs to fit the proposed MAFW dataset.
LOGO-Former (Ma et al., 2023)	2023	DS	Reduce parameter and improve accuracy with modifications in attention.
EST (Liu et al., 2023)	2023	Clip	Propose prediction task to restore unordered fragments to enhance temporal ability.
ESTLNet (Gong et al., 2024)	2024	DS	Designing enhancement methods for occlusion, pose, lighting and temporal issues.

by designing enhancement methods for both spatial and temporal features. Spatial feature enhancement uses multi-level convolution and masked convolution, while temporal feature enhancement uses stacked temporal Transformer blocks and GRUs.

Table 1 lists the contributions of several DFER task models in recent years. DS represents dynamic sampling. TI represents time interpolation. Clip represents clip sampling. Multimodal represents visual, audio and text inputs. The research in Table 1 has contributed extremely brilliant ideas in the DFER. Some of them focus on improving attention mechanisms and Transformer structures, while others design diverse feature enhancement methods. However, few methods address sample defects like occlusion or utilize facial key-points and graph convolution techniques to tackle challenge in DEFR. Thus, we propose a novel network that use facial key-points feature fusion, GCN and temporal attention with relatively low computational cost.

2.2. Self-attention

For temporal information in videos, besides recurrent convolutional networks (Zaremba et al., 2014) and C3D (Tran et al., 2015), Transformer can also explore temporal information well. Transformer (Vaswani et al., 2017) was first proposed in the field of natural language processing, and made a great breakthrough by proposing Transformer for machine translation based on self-attentive mechanism. Girdhar et al. (2019) proposed a video action recognition network based on Transformer that can aggregate spatio-temporal contextual features of video action recognition networks. In 2020, Vision Transformer (ViT) (Dosovitskiy et al., 2020) network was proposed for classification tasks and Swin-Transformer (Liu et al., 2021) was proposed to enhance the network's focus on local regions, which greatly improved the performance of Transformer in computer vision. In the field of DFER, Former-DFER (Zhao and Liu, 2021) sequentially uses spatial attention module and temporal attention module, and utilizes the Transformer architecture to complete the fusion of spatiotemporal features. STT (Ma et al., 2022) integrates spatial attention mechanism and temporal attention mechanism into the same module, immediately executes temporal attention after executing spatial attention. LOGO-FORMER (Ma et al., 2023) reduces parameter and improves expression recognition accuracy through improvements in attention.

2.3. Graph convolution network

Graph convolution network (GCN) is a network used to extract the features of graph structure data. Compared with traditional RNN and CNN, GCN has excellent performance in processing data with unique point and edge structures. Kipf and Welling (2016) used GCN to complete semi-supervised classification tasks. Inspired by this, GCN has attracted more attention in recent years. Many works have raised improvements on the network (Zanfir and Sminchisescu, 2018; Zhao et al., 2019; Liao et al., 2022) and been applied to a variety of tasks. Yan et al. (2018) proposed spatial–temporal GCN to explore key-points tracing on multi-frame dynamic skeleton. Chen et al. (2020) used GCN with abstract scene graphs in cross-modal visual language tasks, which can predict both the importance of different objects in different time steps and the spatial relationship between multiple objects. Wang et al. (2020b) fused the graph matching into GCN to avoid the occlusion and folding of difficult samples. Motivated by this, in the field of DFER, we borrow the rich function of graph convolution, and use GCN to assist in extracting the spatial information of key-points of facial expression, so as to reduce the difficulty of occlusion and side face in DFER datasets.

3. The proposed method

The overall framework of the DFER network proposed in this paper is shown in Fig. 2, which mainly includes a spatial feature extraction network and a temporal feature extraction network, and an effective feature fusion strategy is designed between them. Firstly, the face frames selected from the video are sent to the spatial feature extraction network (FEM) to extract global and local feature vector groups respectively. Then these feature vector groups are sent to the graph convolution network (GCN) for optimization. After that, feature vector groups corresponding to each frame are fused into one through a feature fusion module. Finally, the enhanced spatial information from each frame in the same video are sent to the temporal module to explore the temporal information and perform the final classification prediction.

3.1. Spatial feature extraction network

The spatial feature extraction network mainly consists of two parts: face image feature extraction module (FEM) and graph convolution network (GCN) enhancement module. As shown in Fig. 3, the two networks are cascaded together to extract and optimize global and local information from the input intra-frame image. These two networks will be described in detail below.

3.1.1. Feature extraction module

The facial spatial feature extraction module simultaneously extracts the semantic features of the image and the spatial features of the facial key-points. The upper CNN branch takes a ResNet18 as backbone and make slight modifications on it. We keep the average pooling layer and fully connected layer while change the stride of last two convolution layers to obtain a larger feature map. The triplet attention module (Misra et al., 2021) has been proved to perform well in

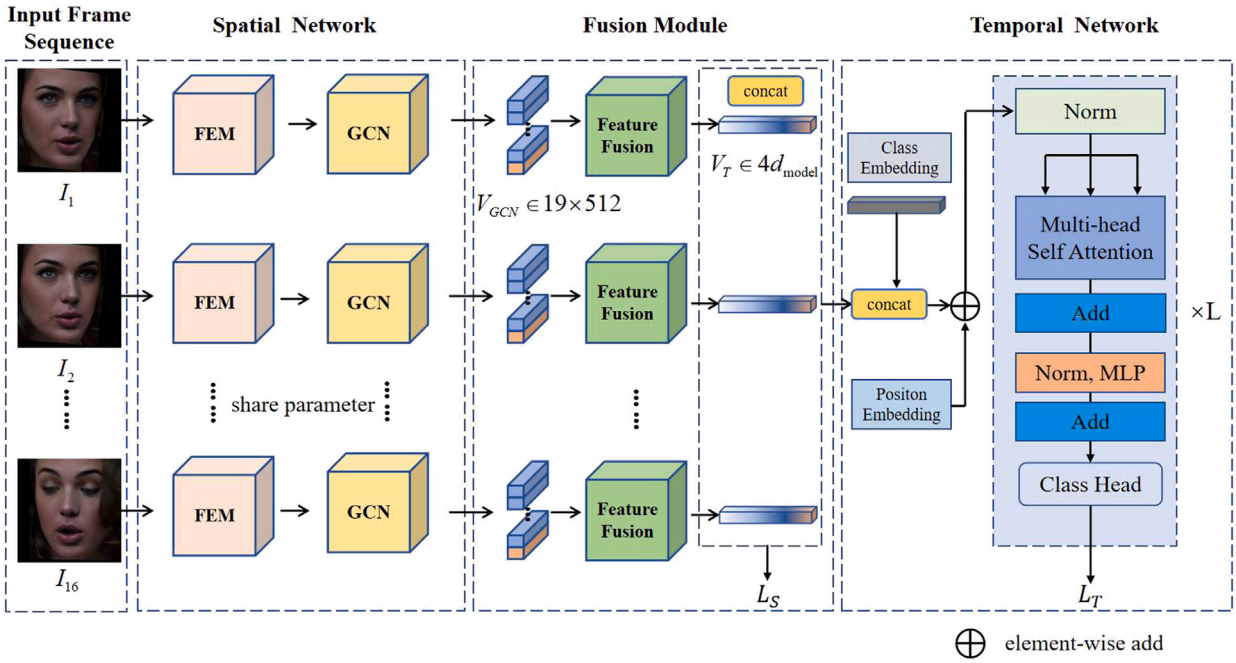


Fig. 2. The overall network architecture.

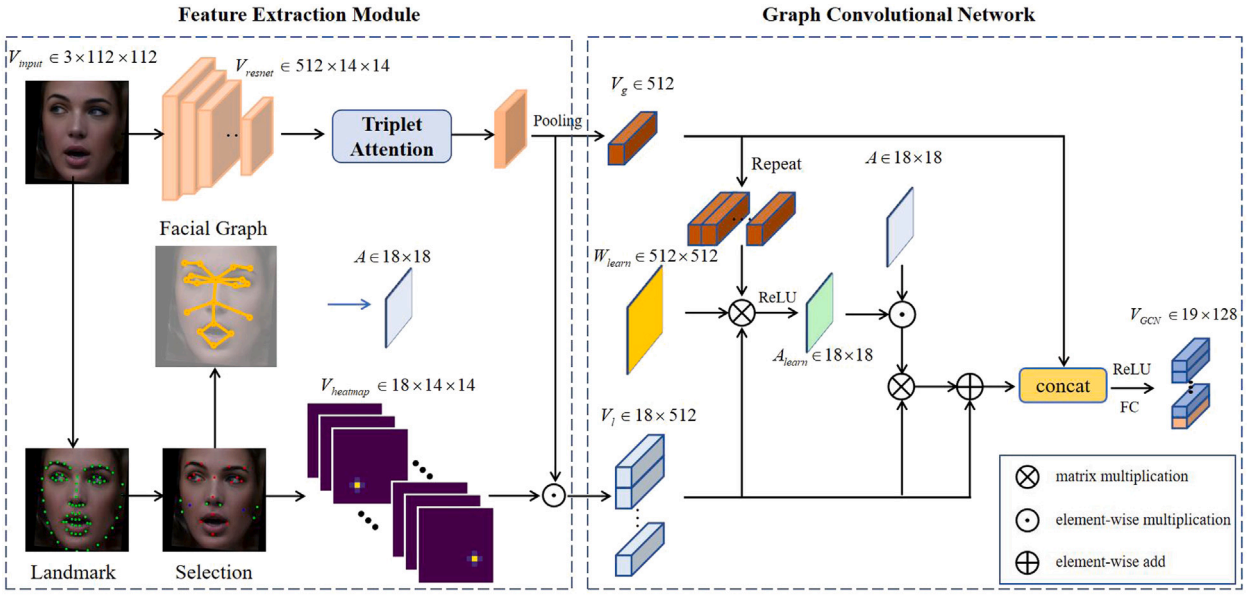


Fig. 3. Diagram of spatial feature extraction network.

processing interactive information of channel dimension and spatial dimension. So, we apply the triplet attention to process feature maps for richer expression information. For a video sample, N frames of face images $X = \{x_1, x_2, \ldots, x_N\}$ are selected, and these images are sent to the spatial feature extraction network respectively. The global feature map F_i is obtained from the input image x_i after ResNet18 and Triplet-Attention. The formula is as follows:

$$F_i = Tri\left(Resnet\left(x_i\right)\right) \tag{1}$$

where $Resnet(\cdot)$ represents the adjusted ResNet18 and $Tri(\cdot)$ represents Triplet-Attention. In the landmark guided attention branch, we utilize dlib to detect 68 face landmarks (the green points on the woman face) from the input face image. We select 16 key-points (the red points) representing eyebrows, eyes, mouth and nose from 68 landmarks based on location. In particular, the cheek part also contains rich expression information, we propose two extra key-points representing the cheek which are calculated from the other neighbor landmarks, shown as blue points. In detail, we choose a fixed three surrounding points to calculate one cheek point, these three points form a triangle area. We regard the center of gravity of triangle areas as the key-points of two cheeks. As shown in Fig. 3, the 18 key-points are obtained and taken as the center to generate 18 Gauss distribution attention heat maps $A_i^j\ (j = 1, 2, \ldots, 18)$. These Gaussian heatmaps are used to guide the feature maps of CNN branches, which are copied to the same depth

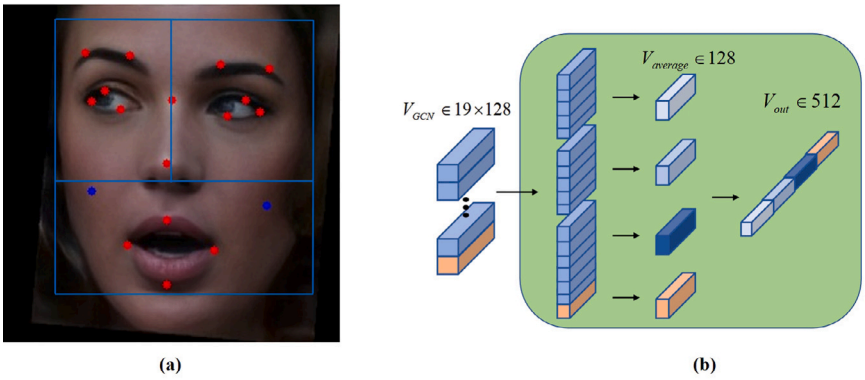


Fig. 4. Diagram of feature fusion module.

as the feature map and multiplicated by the corresponding elements. Finally, a set of local feature vectors is obtained by formula:

$$v_{i,l}^{j} = g\left(F_i \odot A_i^{j}\right) \tag{2}$$

where $\odot$ represents element-wise multiplication and $g(\cdot)$ represents global average pooling. The global average pooling operation is carried out on F_i to obtain the vector $v_{i,g}$ with global information. Finally, the set of vectors output by the feature extraction module is represented by V_i and $V_i = \left\{v_{i,l}^{1}, v_{i,l}^{2}, \ldots, v_{i,l}^{18}, v_{i,g}\right\}$.

3.1.2. Key-points optimization with GCN

Under in-the-wild facial expression datasets, many faces have interference factors such as occlusion, side face, and light shadow. The 18 key-points we obtained may also be affected by these interference factors. To suppress the interfering factors and emphasize the undisturbed local information, we simplified a graph convolutional network from ADGC (Wang et al., 2020b). As shown in Fig. 3 right, the graph convolutional network uses the relationship between the whole and the part to obtain the local information that needs to be emphasized. The formula is as follows:

$$V_i^{GCN} = f[W_a \odot A \otimes V_i^{l} + V_i^{l}]; V_i^{g} \tag{3}$$

$$W_a = ReLU(V_i^{l} \otimes W_{learn} \otimes V_i^{g^T}) \tag{4}$$

where $\otimes$ represents matrix multiplication, $\odot$ represents element-wise multiplication, f represents fully connection layer with $ReLU$ and ; represents concatenation. W_{learn} is used to learn from the input vector V_i^{l} and V_i^{g}. A is the adjacency matrix of 18 key-points as shown in Fig. 2 right. After one residual connection and fully connect layer, the final V_i^{G} is similar to the concatenation of V_i^{l} and V_i^{g} as input but dimension is reduced from 512 to 128. Therefore, the output of a video sample through the feature extraction module and the graph convolution network is $V_{GCN} = \left\{V_1^{G}, V_2^{G}, \ldots, V_N^{G},\right\}$.

3.2. Feature fusion module

In order to explore the temporal relationship between frames in video in the temporal feature extraction network, the feature vector group V_i^{G} corresponding to each frame needs to be fused to obtain a feature vector v_i that can represent a frame of facial image. This paper proposes a feature fusion method based on face spatial structure.

The fusion method based on face spatial structure is to divide the face into several regions, calculate the mean value of local feature vectors in each region, and then splice them. As shown in Fig. 4(a), the face is divided into three areas according to the face structure: the upper left part of the left eye eyebrow, the upper right part of the right eye eyebrow, and the lower part of some cheeks and the whole mouth. Then, as shown in Fig. 4(b), calculate the mean value of the feature vector corresponding to the face key-points in the three regions, regardless of the two key-points on the bridge of the nose, that is, the mean value of the five local feature vectors in the upper left part, the mean value of the five local feature vectors in the upper right part and the mean value of the six local feature vectors in the lower half. Finally, three feature vectors representing various regions are obtained. The three feature vectors are compared with the global eigenvector V_i^{g} to obtain a feature vector v_i^{out} representing the ith frame.

After spatial feature fusion, the features $F_{spatial} = \left\{v_1^{out}, v_2^{out}, \ldots, v_N^{out}\right\}$ of the whole video are obtained. Each feature vector here represents the information of a frame of face image.

In order to make the spatial feature extraction network prefer to obtain expression information and reduce the impact of identity information and interference information on the temporal feature extraction network, an auxiliary loss function is used to supervise the spatial feature extraction network. Firstly, take the average value of the fused video feature $F_{spatial}$, then pass through the full connection layer, and finally calculate the cross-entropy loss. The formula is as follows:

$$L_S = L_{\text{class}}\left(FC\left(\frac{1}{N}\times\sum_{n=1}^{N} v_n\right)\right) \tag{5}$$

where $FC(\cdot)$ is a full connection layer network, and $L_{class}(\cdot)$ is cross entropy loss of multi classification.

3.3. Temporal feature extraction network

The self-attention (Vaswani et al., 2017) model can obtain the temporal information in long-distance sequences well, so this work uses a method based on the Multi-head Self-attention (MSA) module to explore the timing relationship from inter-frames to accurately determine the expression class of the video.

In order to better explore the timing information and obtain the final expression classification results, the video feature $F_{spatial}$ needs to be processed before it is input into the multi-head self-attention module. The formula is as follows:

$$z^{0} = \left[x_{\text{class}}\ ; F_{\text{spatial}}\ \right] + E_{\text{pos}}\ = \left[x_{\text{class}}\ ; v_1; v_2; \cdots; v_N\right] + E_{\text{pos}} \tag{6}$$

where x_{class} is of the same size as v_i. Since the output of the self-attention module is also an N feature vector, which is not conducive to the final expression classification task, a randomly initialized learnable vector x_{class} is used here to implement the classification task. is a learnable position embedding to represent the timing information, and E_{pos} is also randomly initialized.

The query vector q, the key vector k, and the value vector v are then computed for these features with the following equations:

$$\left[q^{(l,h)}, k^{(l,h)}, v^{(l,h)}\right] = LN\left(z^{l-1}\right)\left[W_q^{(l,h)}, W_k^{(l,h)}, W_v^{(l,h)}\right] \tag{7}$$

where $LN(\cdot)$ denotes layer normalization and W is the learnable parameter matrix. In order to learn more possibilities and expand the width of the network, a multi-headed self-attentive structure is used here, so here h represents the hth attention head, $h \in \{1, \ldots, H\}$, and H is the hyperparameter of multi-heads. In addition, a single-layer multi-headed self-attentive network cannot learn the timing information between frames adequately, so a multi-layer serially connected multi-headed self-attentive network is used here, where l represents the lth layer of the multi-headed self-attentive network, $l \in \{1, \ldots, L\}$, and L denotes the number of layers.

Then the self-attentive weights are calculated using the query vector q and the key vector k, and the formula is as follows:

$$A^{(l,h)} = \mathrm{softmax}\left(\frac{q^{(l,h)}\left(k^{(l,h)}\right)^T}{\sqrt{d_k}}\right) \tag{8}$$

where $(\cdot)^T$ denotes the matrix transpose and d_k denotes the dimensions of q and k. The attention matrix $A^{(l,h)}$ is used to optimize the value vector v with the following equation:

$$s^{(l,h)} = A^{(l,h)} v^{(l,h)} \tag{9}$$

Since a multi-headed self-attentive network is used, the values of the different attention heads need to be fused. In addition, a residual structure is used to ensure that no information is lost. The formula is as follows:

$$y^l = W_H \begin{bmatrix} s^{(l,1)} \\ \vdots \\ s^{(l,H)} \end{bmatrix} + z^{l-1} \tag{10}$$

where W_H is the learnable parameter matrix. After fusing the information from multiple attention heads, a multilayer perceptron (MLP) is also used to further optimize the features. The final output of the lth layer multi-headed self-attentive network is obtained as follows:

$$z^l = MLP\left(LN(y^l)\right) + y^l \tag{11}$$

The z^L is obtained after L-layer multi-headed self-attentive network. a learnable classification vector x_{class} has been added in z^0, so the trained classification vector x_{class}^L is obtained at the same position in z^L. x_{class}^L is operated by fully connected layers to obtain the final expression classification vector v_{class}.

The multiclassification cross-entropy loss is calculated for v_{class} with the following equation:

$$L_T = L_{class}\left(v_{class}\right) \tag{12}$$

Therefore, in the training phase, the total loss function of the video expression recognition network proposed in this paper is as follows:

$$L = \lambda \times L_S + (1 - \lambda) \times L_T \tag{13}$$

where λ is a hyperparameter and the value of λ is set to 0.2 by the ablation experiment. in the experimental section of this paper, this ablation experiment will be visualized. In the inference stage, only the final output of the temporal feature extraction network, v_{class}, is used as the expression classification prediction result.

4. Experiments and analysis

4.1. Datasets

To verify the effectiveness of the proposed method, we conduct experiments were conducted on four in-the-wild DFER datasets. The details of the in-the-wild DFER datasets are summaried in Table 2.

Table 2
The details of the in-the-wild DFER datasets.

Dataset	Training set	Validation set	Fold nums
AFEW (Dhall et al., 2018)	773	383	One-fold
DFEW (Jiang et al., 2020)	≈ 9400	≈ 2300	Five-fold
FERV39K (Wang et al., 2022a)	31 088	7847	One-fold
MAFW (Liu et al., 2022a)	≈ 7300	≈1830	Five-fold

AFEW (Acted Facial Expressions In The Wild) (Dhall et al., 2018) is an in-the-wild dataset with a collection of video clips from different film and television productions. The AFEW has been the evaluation dataset for the Emotion Recognition in The Wild Challenge (EmotiW) from 2013 to 2019, during which time the dataset was updated. Expressions in the AFEW dataset are spontaneous and the AFEW is a temporal multimodal database containing both audio and video data. The samples are labeled with seven expressions and 1809 videos are available in the AFEW dataset, including 773 in the training set, 383 in the validation set, and 653 in the test set. To ensure data rigor, there are no duplicate videos in these three sets, and even the identities of the people in the videos are not the same. Since the test set is not publicly available, this paper uses the training set and the validation set for experiments.

DFEW (Dynamic Facial Expressions in the Wild) (Jiang et al., 2020) is an in-the-wild dataset proposed in 2020. The dataset contains samples from more than 1500 close-to-life HD movies covering a variety of topics that realistically reflect people's facial movements in various environments. The final dataset production team edited 16,372 video samples, and a total of 12 experts annotated each video 10 times independently. These samples were classified into seven basic expressions. In addition, the dataset production team extracted image frames for each video and removed the background information from each frame, which greatly reduced the cost of dataset pre-processing. The data are evaluated using a five-fold cross-validation, where the samples of this dataset are evenly divided into five non-repeating parts, and one of the parts is selected as the test set and the other samples are used as the training set.

FERV39K (Wang et al., 2022a) dataset is a DFER dataset publicly released by a research team from Fudan University in 2022. It consists of 38935 video segments and 7 emotion categories, making it the largest dynamic expression recognition dataset currently available. Each video sample is independently annotated by 30 professional annotators to ensure high-quality labels are obtained. This dataset is divided into four categories according to different scene environments, including Daily Life, Week interactive shows, Strong interactive shows, and Anomaly Issues. It can also be further subdivided into 22 more specific scenes. Thanks to the large sample size, each of the four main subsets contains about 10000 video samples. Typically, on this dataset, 80% is used as the training set and 20% as the validation set to partition and obtain model evaluation metrics.

MAFW (Liu et al., 2022a) dataset is a large-scale in-the-wild facial DFER dataset publicly released by China University of Geosciences in 2022. This dataset contains 10045 video samples and is the first large-scale multimodal emotion recognition task dataset with single category labels, multi category labels, and Chinese English sentiment description text labels. When recognizing a single expression, these samples are divided into 11 categories, including anger (AN), disgust (DI), fear (FE), happiness (HA), sadness (SA), surprise (SU), contempt (CO), anxiety (AX), helplessness (HL). Disappointment (DS), Neutrality (NE). Similar to the DFEW dataset, it divides into five subsets for five fold cross validation.

4.2. Implementation details

Our method is implemented with Pytorch framework and trained on a Titan GPU with 24 GB of memory. During the training phase, the size of each batch is set to 64 and epoch is set to 80. The initial learning

Table 3
Comparison with SOTA methods on DFEW. The best results are highlighted in **bold**, the suboptimal result are highlighted in underline. The indicators in the table are the average results of the DFEW five-fold cross validation experiment.

Method	Params (M)	Accuracy of 7 emotion classes							Metric	
		Happy	Sad	Neutral	Anger	Surprise	Disgust	Fear	UAR (%)	WAR (%)
C3D (Tran et al., 2015)	78	75.17	39.49	55.11	62.49	45.00	1.38	20.51	42.74	53.54
P3D (Qiu et al., 2017)	98	74.85	43.40	54.18	60.42	50.99	0.69	23.28	43.97	54.47
3D Resnet18 (Hara et al., 2018)	33	76.32	50.21	64.18	62.85	47.52	0.00	24.56	46.52	58.27
Resnet18+LSTM (Zhao and Liu, 2021)	–	83.56	61.56	68.27	65.29	51.26	0.00	29.34	51.32	51.32
Resnet18+GRU (Zhao and Liu, 2021)	–	82.87	63.83	65.06	68.51	52.00	0.86	30.14	51.68	64.02
Former-DFER (Zhao and Liu, 2021)	18	84.05	62.57	67.52	70.03	**56.43**	3.45	31.78	53.69	65.70
STT (Ma et al., 2022)	–	87.36	67.90	64.97	71.24	53.10	3.49	34.04	54.58	66.45
CEFLNet (Liu et al., 2022b)	**13**	84.00	**68.00**	67.00	70.00	52.00	0.00	17.00	51.14	65.35
NR-DFERNet (Li et al., 2022)	19	86.42	65.10	70.40	**72.88**	50.10	0.00	**45.44**	55.77	68.01
DPCNet (Wang et al., 2022b)	51	**89.93**	64.61	67.12	63.18	53.67	**15.86**	31.56	55.13	66.32
EST (Liu et al., 2023)	43	86.87	66.58	67.18	71.84	47.52	5.52	28.49	53.43	65.85
LOGO-Former (Ma et al., 2023)	–	85.39	66.52	68.94	71.33	54.59	0.00	32.71	54.21	66.98
Ours	17	88.42	67.39	**71.75**	72.72	54.77	0.00	36.90	**55.85**	**68.73**

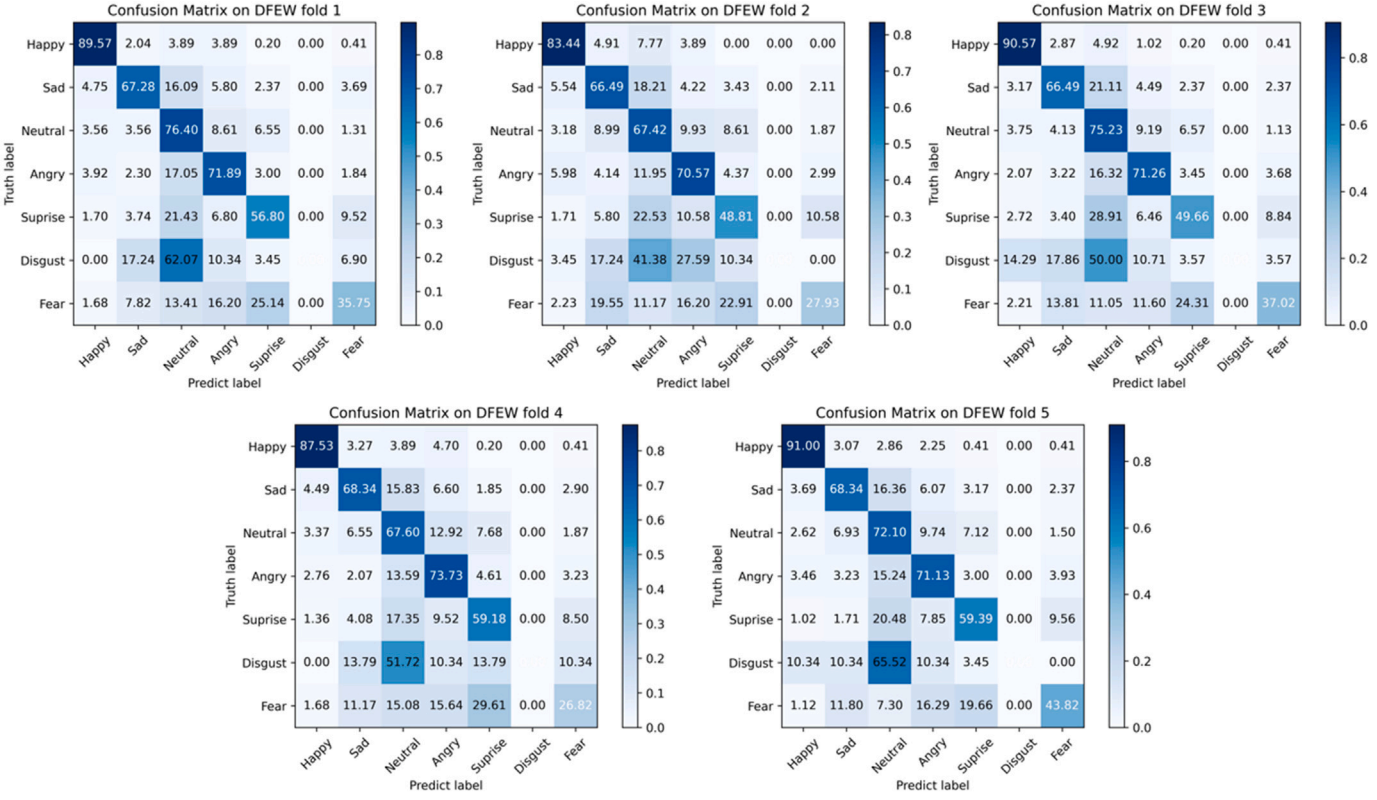


Fig. 5. Confusion Matrix on DFEW dataset.

rate is 0.001 and is multiplied by 0.1 at the 40th and 60th cycles. Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ is adopted. For the multi-headed self-attentive module, let $H = 8$, which is the 8-headed self-attentive module. In addition, since the expression features have been extracted in the spatial feature extraction module and the self-attentive module is mainly used to explore the temporal information and classification, only a 3-layer multi-headed self-attentive network is used, $L = 3$.

4.3. Comparison to state of the art models

Experimental results on four datasets are presented below, which contains UAR and WAR indicators and visualizations.

4.3.1. Comparison on DFEW

As shown in Table 3, our method achieves 68.73% WAR and 55.85% UAR on the DFEW dataset, which is 0.72% higher than NR-DFERNet (Li et al., 2022) on WAR and 0.09% higher on UAR, while reducing the number of parameters by 2M. On the accuracy of the 7 emotional base categories, neutral reached the best level, while the five categories of happiness, sadness, anger, surprise, and fear reached the second best level.

Fig. 5 shows the confusion matrix obtained in the five-fold cross validation experiment on the DFEW dataset. It can be seen from the figure that our method has ideal recognition accuracy in the four categories of happiness, sadness, neutrality and anger, while accuracy on disgust category is 0 like other methods. We argue that the reason is

Table 4
Comparison with SOTA methods on AFEW. The best results are highlighted in **bold**, the suboptimal result are highlighted in underline.

Method	Params (M)	UAR (%)	WAR (%)
EmotiW-Baseline (Dhall et al., 2018)	–	–	38.81
C3D (Tran et al., 2015)	78	43.75	46.72
DenseNet161 (Liu et al., 2018)	27	–	51.44
Emotion-FAN (Meng et al., 2019b)	–	–	51.18
Emotion-BEEU (Kumar et al., 2020)	–	–	52.49
3D ResNet18 (Hara et al., 2018)	33	42.14	45.67
ResNet18+LSTM (Zhao and Liu, 2021)	–	43.96	48.82
ResNet18+GRU (Zhao and Liu, 2021)	–	45.12	49.34
Former-DFER (Zhao and Liu, 2021)	18	47.42	50.92
CEFLNet (Liu et al., 2022b)	**13**	48.29	53.98
EST (Liu et al., 2023)	43	49.57	54.26
ESTLNet (Gong et al., 2024)	–	–	53.79
OURS	17	**50.14**	**55.00**

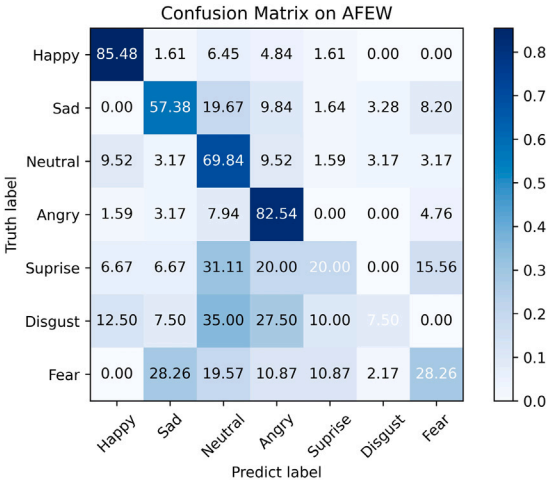


Fig. 6. Confusion Matrix on AFEW dataset.

the serious class imbalancement in the dataset, with less than 2% of the samples being disgusted. At the same time, disgust itself has ambiguity, and people usually suppress disgust in natural scenes. Therefore, its facial features are similar to neutrality, leading to a decrease in the accuracy of the model's recognition of disgust categories.

4.3.2. Comparison on AFEW

As shown in Table 4, our method achieves 55.00% WAR and 50.14% UAR on the AFEW dataset. Compared to the EST (Liu et al., 2023) model, ours improves UAR by 0.57%, WAR by 0.74%, and reduces parameter quantity by 60.5%. The overall results are significantly better than other methods. Fig. 6 shows the confusion matrix calculated on the AFEW dataset. Results are similar to those on DFEW dataset, with ideal performance in the four categories of happiness, sadness, neutrality, and anger, while the other three categories perform poorly. On the one hand, it is because the AFEW dataset also has class imbalance issues. On the other hand, it is due to the fine-tuning of the model trained from DFEW, which results in poor results in three categories.

4.3.3. Comparison on FERV39k

The results of our method on the FERV39k dataset are shown in Table 5. Our method perform well on neutral and anger expressions and achieved 47.80% WAR and 35.16% UAR on FERV39K, slightly weaker than the STT (Ma et al., 2022), NR-DFERNet (Li et al., 2022) and LOGO-Former (Ma et al., 2023) models.

Thanks to the large amount of data and detailed scene segmentation of FERV39k, researchers can explore strategies to improve the accuracy of expression recognition under different scene data. Table 6 shows the performance indicators of our method on four main subsets of FERV39k (DL11k, WIS9k, SIA10k and AI9k) and several sub subsets (social, conflict, Argue, Action and ElegantArt), with significant improvements compared to several baseline methods (Wang et al., 2022a) and their improved methods.

4.3.4. Comparison on MAFW

The results on the MAFW dataset are shown in Table 7. The overall performance of our method is high, achieving 47.44% WAR and 33.39% UAR. Compared with the T-ESFL (Liu et al., 2022a) model, it lags behind by 0.74% in WAR and exceeds 0.11% in UAR, and is significantly higher than other models. Our method achieved the best results in sadness (SA) and anxiety (AX), followed by the second best results in anger (AN) and helplessness (HL) categories, and is similar to the best results in other categories.

4.3.5. Model size and inference speed

Table 8 compares parameter quantity and inference efficiency of some methods and our model. Parameter quantity of some other models have been shown in the previous tables. To make a fair comparison with these models, we use the full model to measure the computational cost and conduct experiments on a RTX3090 GPU. The training time is measured on the first-fold of DFEW dataset of 80 epoches even though our model achieved the best results in 65th epoch, Former-DFER reach the best in 93th epoch, and NR-DFER reach the best in 90th. Our method results in the competitive performance with a considerable cost and processing speed. It shows the possibility of our method being applied since our model can process real-time input with a sequence of 16 frames.

4.4. Ablation experiment

4.4.1. Effectiveness of the architecture design

The method proposed in this paper contains different modules, and to verify the impact of these modules on DFER, ablation experiments were designed on the DFEW dataset as in Table 9. The different structures are shown in Fig. 7 and multi-frame legend is omitted in order to simplify the drawing. Fig. 7(a) shows the baseline approach for the whole network, using only the feature extraction module (FEM) in the spatial feature extraction network to extract features, and then calculating the mean value of all feature vectors for all frames and using the fully connected layer to obtain the classification vectors without using the graph convolution network and the temporal feature extraction network; Fig. 7(b) only uses the spatial feature extraction network; Fig. 7(c) uses LSTM instead of the multi-headed self-attentive model MSA to resolve the differences between temporal networks; Fig. 7(d) is designed to verify the effect of the graph convolution module on video face expression recognition; Fig. 7(e) is the full network proposed in this paper.

As can be seen from Model A and Model B in Table 9, the addition of the graph convolution optimization module can improve the recognition accuracy by 0.68% over the baseline method. Also for the ablation experiment of the graph convolution module, Model E improves the recognition accuracy by 1.43% over Model D, which indicates that the self-attentive module in the temporal feature extraction network can further amplify the performance of the graph convolution module. The recognition accuracy of model E is 3.25% higher than that of model B, which illustrates the importance of the self-attentive module for video face expression recognition. And the comparison results of model C and model E show that the self-attentive module is more capable of acquiring temporal expression information in videos than the LSTM.

Table 5
Comparison with SOTA methods on FERV39k. The best results are highlighted in **bold**, the suboptimal result are highlighted in underline.

Method	Params (M)	Accuracy of 7 emotion classes							Metric	
		Happy	Sad	Neutral	Anger	Surprise	Disgust	Fear	UAR (%)	WAR (%)
C3D (Tran et al., 2015)	78	48.20	35.53	52.71	13.72	3.45	4.93	0.23	22.68	31.69
P3D (Qiu et al., 2017)	98	61.85	42.21	49.80	42.57	10.50	0.86	5.57	30.48	40.81
3D Resnet18 (Hara et al., 2018)	33	57.64	28.21	59.60	33.29	4.70	0.21	3.02	26.67	37.57
Resnet18+LSTM (Zhao and Liu, 2021)	–	61.91	31.95	61.70	45.93	14.26	0.00	0.70	30.92	42.59
Former-DFER (Zhao and Liu, 2021)	18	65.65	51.33	56.74	43.64	**21.94**	8.57	**12.53**	37.20	46.85
STT (Ma et al., 2022)	–	**69.77**	47.81	59.14	47.41	20.22	**10.49**	9.51	37.76	48.11
NR-DFERNet (Li et al., 2022)	19	69.18	**54.77**	51.12	49.70	13.17	0.00	0.23	35.82	**48.54**
LOGO-Former (Ma et al., 2023)	–	–	–	–	–	–	–	–	**38.22**	48.13
Ours	17	64.90	49.95	**66.24**	**51.54**	13.25	0.00	0.00	35.16	47.80

Table 6
Comparison with menchmarks on FERV39k subsets. The best results are highlighted in **bold**, the suboptimal result are highlighted in underline.

Method	All	DL11k	WIS9k	SIA10k	AI9k	Social	Conflit	Argue	Action	Art
R18	39.33	39.75	40.50	42.31	33.90	39.74	39.52	44.09	50.61	33.33
R50	30.57	30.46	32.52	30.56	30.14	27.51	31.14	36.96	37.80	31.35
C3D	31.69	26.95	30.15	42.70	27.29	34.50	21.96	31.52	35.98	28.97
I3D	38.78	38.56	38.52	40.55	37.44	37.55	34.93	43.34	39.63	37.72
Ours	**47.80**	**49.88**	**44.83**	**46.56**	**47.65**	**52.73**	**47.66**	**59.28**	**59.46**	**55.74**

Table 7
Comparison with SOTA methods on MAFW. The best results are highlighted in **bold**, the suboptimal result are highlighted in underline. The indicators in the table are the average results of the MAFW five-fold cross validation experiment.

Method	Accuracy of 11 emotion classes											Metrics	
	AN	DI	FE	HA	NE	SA	SU	CO	AX	HL	DS	UAR(%)	WAR (%)
ResNet18 (He et al., 2016)	45.02	9.25	22.51	70.69	35.94	52.25	39.04	0.00	6.67	0.00	0.00	25.58	36.65
ViT (Dosovitskiy et al., 2020)	46.03	**18.18**	27.49	76.89	50.70	68.19	45.13	1.27	18.93	1.53	1.65	32.36	45.04
C3D (Tran et al., 2015)	51.47	10.66	24.66	70.64	43.81	55.04	46.61	1.68	24.34	**5.73**	**4.93**	31.17	42.25
Res+LSTM (Zhao and Liu, 2021)	46.25	4.70	25.56	68.92	44.99	51.91	45.88	**1.69**	15.75	1.53	1.65	28.08	39.38
ViT+LSTM (Liu et al., 2022a)	42.42	14.58	**35.69**	76.25	54.48	68.87	41.01	0.00	24.40	0.00	1.65	32.67	45.56
C3D+LSTM (Liu et al., 2022a)	54.91	0.47	9.00	73.43	41.39	64.92	**58.43**	0.00	24.62	0.00	0.00	29.75	43.76
T-ESFL (Liu et al., 2022a)	**62.70**	2.51	29.90	**83.82**	**61.16**	67.98	48.50	0.00	9.52	0.00	0.00	33.28	**48.18**
Ours	61.87	13.28	15.20	74.40	51.75	**69.18**	47.20	0.00	**28.42**	5.66	0.00	**33.39**	47.44

Table 8
Comparison on parameter quantity and inference efficiency.

Method	Params (M)	Inference speed (samples/s)	Training time (h)	WAR on DFEW
Dense-161 (Liu et al., 2018)	27	4	19	60.40
Former-DFER (Zhao and Liu, 2021)	**13**	**86**	17.5	65.70
NR-DFER (Li et al., 2022)	19	76	8.5	68.01
Ours	17	52	**6**	**68.73**

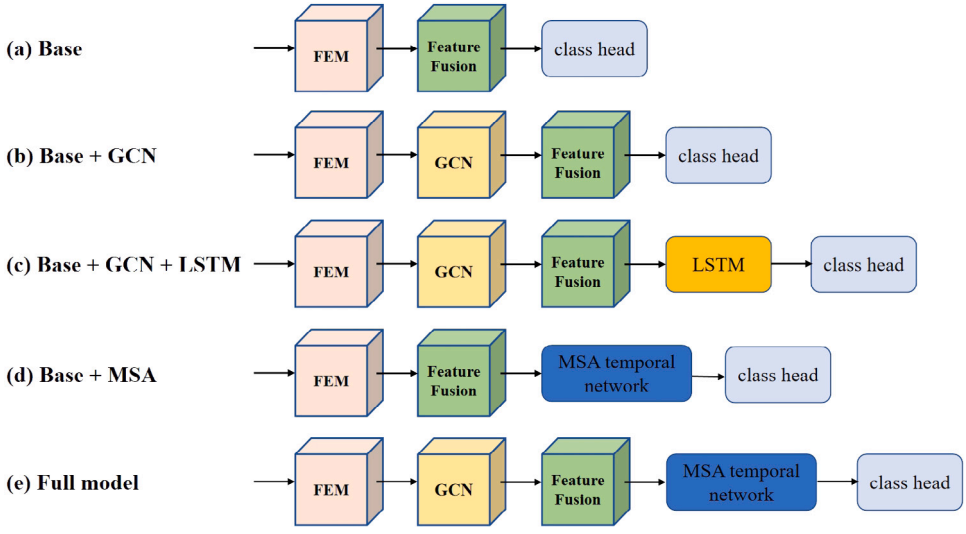


Fig. 7. Ablation Study on proposed components.

Table 9
Ablation study on proposed components.

Model	FEM	GCN	LSTM	MSA	WAR(%)
A	√	×	×	×	64.80
B	√	√	×	×	65.48
C	√	√	√	×	66.72
D	√	×	×	√	67.30
E	√	√	×	√	68.73

Table 10
Ablation experiment of spatial feature fusion module.

Method	DFEW WAR(%)	AFEW WAR(%)
mean-based fusion method	67.07	53.95
LSTM-based fusion method	67.02	53.68
structure-based fusion method	68.73	**55.00**

4.4.2. Ablation study on fusion module

To verify the effect of spatial feature fusion module, we designed an ablation experiments with other fusion methods in Table 10. The mean-based fusion method is to directly calculate the mean value for the feature vector group. The LSTM-based fusion method is to input the feature vector group into the LSTM network, and the output of the last node is used as the fusion result. From the experimental results in Table 10, we can see that the designed fusion method based on face spatial structure can achieve better expression recognition results, which shows that the study of face spatial structure is beneficial to the research work of expression recognition.

4.4.3. Ablation study on hyperparameters

The total loss function of our method is $L = \lambda \times L_S + (1 - \lambda) \times L_T$. In order to obtain a better weight of the loss function, we designed experiments for analysis, and the experimental results are shown in Fig. 8. Since L_S is an auxiliary loss function and L_T is the final classification loss, λ cannot exceed 0.5. From the experimental results in the figure, we can see that the best expression recognition result of 68.73% can be achieved when $\lambda = 0.2$.

We also conduct experiments to verify the effectiveness of temporal transformer block depth and Multi-head Self-attention (MSA) module on the accuracy of expression recognition. As shown in Tables 11 and 12, the best result can be achieved when heads of Self-attention equal 8 and depth of transformer blocks equals 3. Since the spatial feature has been extracted, temporal information requires relatively shallow blocks and network with deeper structure performs no better than the shallower one. Moreover, for in-the-wild DFER tasks, deeper network is likely to entail overfitting since the strong class imbalance. We argue that DFER tasks involve video samples with small action amplitudes and a concentration of emotional information. If the stacking depth of the temporal network is too deep, it will blur the already small emotional transition features and pay more attention to all video frame features. Due to the output features of the feature aggregation module are concatenated with four parts representing different facial region features, it is more reasonable to focus attention within the self region features. Therefore, in the experiment, it is found that the performance is significantly better when the number of attention heads is a multiple of 4, such as 8 and 12, compared to other situations.

4.5. Visualization

In addition to the confusion matrix shown in the previous text, t-SNE graphs are commonly used in classification tasks to display the discrimination of each category. Fig. 9 shows the t-SNE plots of Former-DFER, NR-DFER and our method on the DFEW datasets. Fig. 10 shows

Table 11
Comparative experiment of MSA module.

Num of heads	DFEW WAR(%)	AFEW WAR(%)
6	68.01	53.42
8	**68.73**	**55.00**
10	65.70	53.95
12	68.35	54.47

Table 12
Evaluation on different depth of transformer blocks.

Temporal layers	DFEW WAR(%)	AFEW WAR(%)
2	63.56	53.16
3	**68.73**	**55.00**
4	67.41	53.68
5	67.24	51.58
6	67.66	52.11

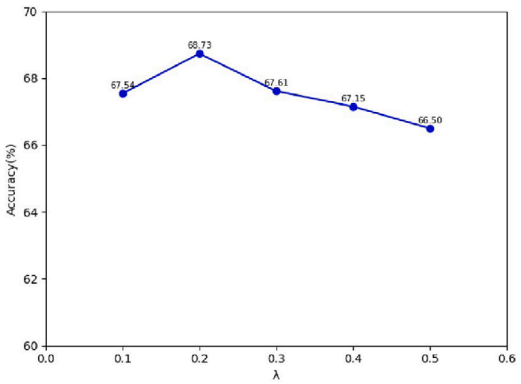


Fig. 8. Ablation Study on loss hyperparameter.

the t-SNE plots of our method on the AFEW, FERV39k and MAFW datasets. Due to significant differences in data volume, we select the first fold validation set of DFEW and MAFW, all samples in AFEW and 2000 samples randomly sampled from the FERV39k dataset.

The t-SNE diagrams of DFEW, AFEW and FERV39k show that the model performs well in four categories: neutral, sad, efficient, and surprised. It can effectively aggregate intra class samples, separate samples from different categories, and mix the sample points of the other three categories together. The t-SNE map of the 11 categories in MAFW has a lower clustering degree than the dataset of the 7 categories, but can still observe category boundaries of happiness, anxiety, sadness, surprise, and helplessness. The above visualization results demonstrate the excellent performance of our method in DFER tasks.

5. Conclusion

A new network for dynamic facial expression recognition (DFER) in-the-wild is proposed in this paper, which is based on spatial key-points optimized region feature fusion and temporal self-attention. The intra-frame spatial expression information is extracted with a facial feature extraction module and optimized by a key-points guided graph convolution module. A face structure-based spatial feature fusion module is designed to fuse the spatial information. In the temporal feature extraction network, a multi-headed self-attention network is used to obtain the temporal information from inter-frames and generate the

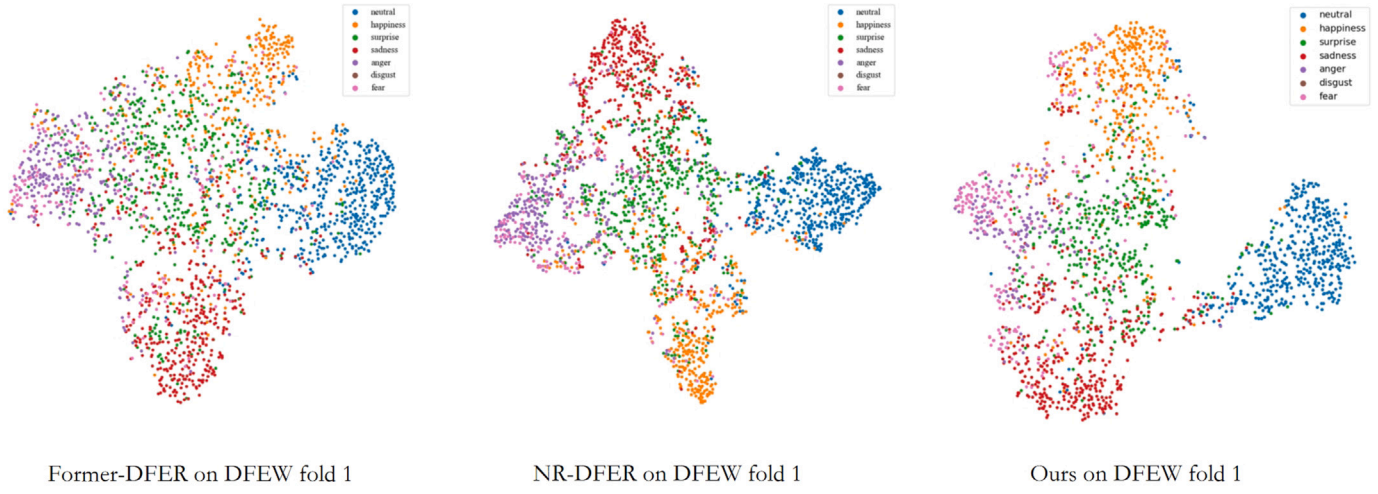


Fig. 9. Comparison of t-SNE on DFEW fold 1.

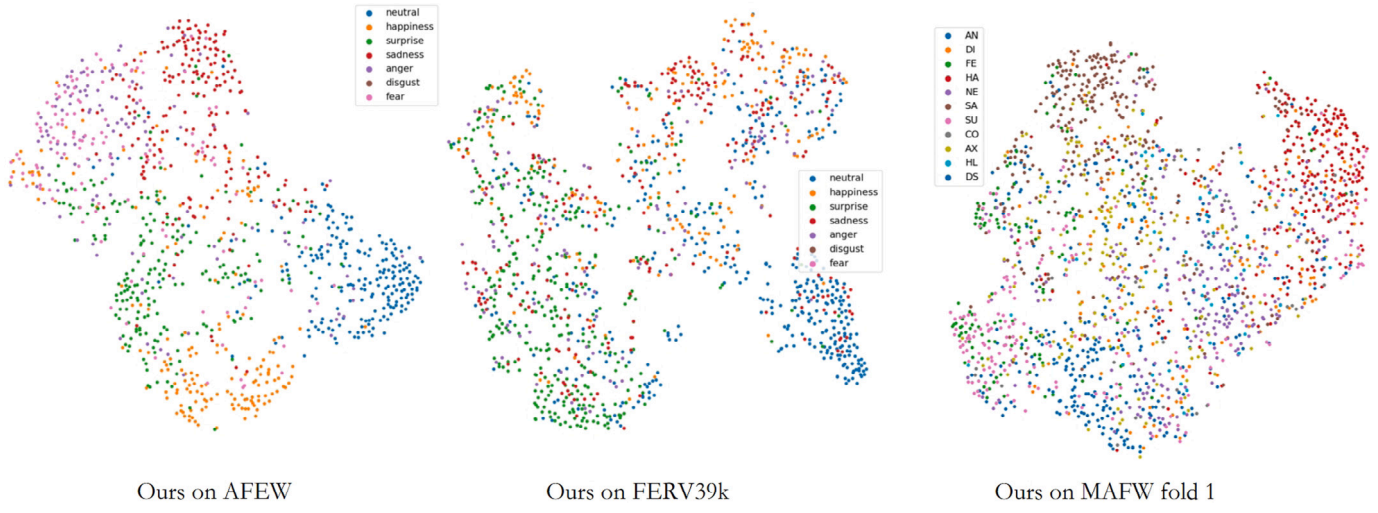


Fig. 10. t-SNE plots of three datasets.

final classification vector. Our method achieves competitive WAR of 55.00% on AFEW, 68.73% on DFER, 47.80% on FERV39k and 47.44% on MAFW. Ablation experiments showed that the GCN module, fusion module, and temporal module improved the accuracy on DFEW by 0.68%, 1.66%, and 3.25%, respectively, which strongly proves the effectiveness of our modules. We also found that shallow temporal network depth in DFER tasks is beneficial for the network to fully utilize its performance. Our method also perform well in terms of parameter quantity and inference speed, illustrating the effectiveness for DFER in-the-wild.

CRediT authorship contribution statement

Zhiwei Huang: Conceptualization, Writing – original draft, Writing – review & editing. **Yu Zhu:** Project administration, Resources, Writing – review & editing. **Hangyu Li:** Validation, Visualization. **Dawei Yang:** Funding acquisition, Supervision.

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.

Data availability

Data will be made available on request.

Acknowledgments

All authors discuss the results and contribute to the final manuscript.

This work is supported in part by the National Natural Science Foundation of China under Grant 82170110, and the Science and Technology Commission of Shanghai Municipality under Grant 20DZ22544000, 21DZ2200600, 20DZ2261200. Fujian Province Department of Science and Technology (2022D014).

References

Baddar, W.J., Ro, Y.M., 2019. Mode variational lstm robust to unseen modes of variation: Application to facial expression recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, (01), pp. 3215–3223.

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

Chen, J., Chen, Z., Chi, Z., Fu, H., 2014. Emotion recognition in the wild with feature fusion and multiple kernel learning. In: Proceedings of the 16th International Conference on Multimodal Interaction. pp. 508–513.

Chen, S., Jin, Q., Wang, P., Wu, Q., 2020. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9962–9971.

Dhall, A., Goecke, R., Joshi, J., Wagner, M., Gedeon, T., 2013. Emotion recognition in the wild challenge 2013. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction. pp. 509–516.

Dhall, A., Kaur, A., Goecke, R., Gedeon, T., 2018. Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. pp. 653–656.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R., Pal, C., 2015. Recurrent neural networks for emotion recognition in video. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. pp. 467–474.

Girdhar, R., Carreira, J., Doersch, C., Zisserman, A., 2019. Video action transformer network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 244–253.

Gong, W., Qian, Y., Zhou, W., Leng, H., 2024. Enhanced spatial-temporal learning network for dynamic facial expression recognition. Biomed. Signal Process. Control 88, 105316.

Hara, K., Kataoka, H., Satoh, Y., 2018. Can spatiotemporal 3D cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6546–6555.

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

Huang, X., He, Q., Hong, X., Zhao, G., Pietikainen, M., 2014. Improved spatiotemporal local monogenic binary pattern for emotion recognition in the wild. In: Proceedings of the 16th International Conference on Multimodal Interaction. pp. 514–520.

Jiang, X., Zong, Y., Zheng, W., Tang, C., Xia, W., Lu, C., Liu, J., 2020. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2881–2889.

Kahou, S.E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç., Memisevic, R., Vincent, P., Courville, A., Bengio, Y., Ferrari, R.C., et al., 2013. Combining modality specific deep neural networks for emotion recognition in video. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction. pp. 543–550.

Kim, D.H., Baddar, W.J., Jang, J., Ro, Y.M., 2017. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. IEEE Trans. Affect. Comput. 10 (2), 223–236.

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Kossaifi, J., Toisoul, A., Bulat, A., Panagakis, Y., Hospedales, T.M., Pantic, M., 2020. Factorized higher-order cnns with an application to spatio-temporal emotion estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6060–6069.

Kumar, V., Rao, S., Yu, L., 2020. Noisy student training using body language dataset improves facial expression recognition. In: European Conference on Computer Vision. Springer International Publishing Cham, pp. 756–773.

Lee, M.K., Choi, D.Y., Kim, D.H., Song, B.C., 2019. Visual scene-aware hybrid neural network architecture for video-based facial expression recognition. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition. (FG 2019), IEEE, pp. 1–8.

Li, H., Sui, M., Zhu, Z., et al., 2022. Nr-dfernet: Noise-robust network for dynamic facial expression recognition. arXiv preprint arXiv:2206.04975.

Liao, L., Zhu, Y., Zheng, B., Jiang, X., Lin, J., 2022. FERGCN: facial expression recognition based on graph convolution network. Mach. Vis. Appl. 33 (3), 40.

Liu, Y., Dai, W., Feng, C., Wang, W., Yin, G., Zeng, J., Shan, S., 2022a. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 24–32.

Liu, Y., Feng, C., Yuan, X., Zhou, L., Wang, W., Qin, J., Luo, Z., 2022b. Clip-aware expressive feature learning for video-based facial expression recognition. Inform. Sci. 598, 182–195.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Liu, C., Tang, T., Lv, K., Wang, M., 2018. Multi-feature based emotion recognition for video clips. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. pp. 630–634.

Liu, Y., Wang, W., Feng, C., Zhang, H., Chen, Z., Zhan, Y., 2023. Expression snippet transformer for robust video-based facial expression recognition. Pattern Recognit. 138, 109368.

Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I., 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 Ieee Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. IEEE, pp. 94–101.

Ma, F., Sun, B., Li, S., 2022. Spatio-temporal transformer for dynamic facial expression recognition in the wild. arXiv preprint arXiv:2205.04749.

Ma, F., Sun, B., Li, S., 2023. Logo-former: Local-global spatio-temporal transformer for dynamic facial expression recognition. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 1–5.

Meng, D., Peng, X., Wang, K., Qiao, Y., 2019a. Frame attention networks for facial expression recognition in videos. In: 2019 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 3866–3870.

Meng, D., Peng, X., Wang, K., Qiao, Y., 2019b. Frame attention networks for facial expression recognition in videos. In: 2019 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 3866–3870.

Misra, D., Nalamada, T., Arasanipalai, A.U., Hou, Q., 2021. Rotate to attend: Convolutional triplet attention module. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3139–3148.

Pantic, M., Valstar, M., Rademaker, R., Maat, L., 2005. Web-based database for facial expression analysis. In: 2005 IEEE International Conference on Multimedia and Expo. IEEE, pp. 5–pp.

Qiu, Z., Yao, T., Mei, T., 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5533–5541.

Taini, M., Zhao, G., Li, S.Z., Pietikainen, M., 2008. Facial expression recognition from near-infrared video sequences. In: 2008 19th International Conference on Pattern Recognition. IEEE, pp. 1–4.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4489–4497.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30.

Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y., 2020a. Suppressing uncertainties for large-scale facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6897–6906.

Wang, Y., Sun, Y., Huang, Y., Liu, Z., Gao, S., Zhang, W., Ge, W., Zhang, W., 2022a. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20922–20931.

Wang, Y., Sun, Y., Song, W., Gao, S., Huang, Y., Chen, Z., Ge, W., Zhang, W., 2022b. Dpcnet: Dual path multi-excitation collaborative network for facial expression representation learning in videos. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 101–110.

Wang, G., Yang, S., Liu, H., Wang, Z., Yang, Y., Wang, S., Yu, G., Zhou, E., Sun, J., 2020b. High-order information matters: Learning relation and topology for occluded person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6449–6458.

Yan, S., Xiong, Y., Lin, D., 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, (1).

Zanfir, A., Sminchisescu, C., 2018. Deep learning of graph matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2684–2693.

Zaremba, W., Sutskever, I., Vinyals, O., 2014. Recurrent Neural Network Regularization. Cornell University, arXiv, Cornell University - arXiv.

Zhao, X., Liang, X., Liu, L., Li, T., Han, Y., Vasconcelos, N., Yan, S., 2016. Peak-piloted deep network for facial expression recognition. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part II 14. Springer International Publishing, pp. 425–442.

Zhao, Z., Liu, Q., 2021. Former-dfer: Dynamic facial expression recognition transformer. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1553–1561.

Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N., 2019. Semantic graph convolutional networks for 3d human pose regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3425–3435.

Zhao, G., Pietikainen, M., 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. 29 (6), 915–928.

Zheng, C., Mendieta, M., Chen, C., 2023. Poster: A pyramid cross-fusion transformer network for facial expression recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3146–3155.