Dynamic-Static Graph Convolutional Network for Video-Based Facial Expression Recognition

Fahong Wang[1], Zhao Liu[2], Jie Lei[1](✉), Zeyu Zou[2], Wentao Han[2], Juan Xu[2], Xuan Li[2], Zunlei Feng[3], and Ronghua Liang[1]

[1] College of Computer Science, Zhejiang University of Technology, Hangzhou, China
{fhwang,jasonlei,rhliang}@zjut.edu.cn

[2] Ping an Life Insurance of China, Ltd., Shanghai, China
{liuzhao556,zouzeyu313,xujuan635,lixuan208}@pingan.com.cn,
wthan@zjut.edu.cn

[3] College of of Computer Science, Zhejiang University, Hangzhou, China
zunleifeng@zju.edu.cn

Abstract. Most of the current methods for video-based facial expression recognition (FER) in the wild are based on deep neural networks with attention mechanism to capture the relationships between frames. However, these methods suffer from the large variations of expression patterns and data uncertainties. This paper proposes a Dynamic-Static Graph Convolutional Network (DSGCN), which mainly consists of a Static-Relational graph (SRG) and a Dynamic-Relational graph (DRG). The SRG aims to guide the network to learn the static spatial relationship of facial expressions in each video frame, strengthening the salient areas of the face through the dependencies of context nodes. The DRG learns the dynamic temporal relationship of facial expressions by aggregating video sequence features, constructing a graph with other samples within a batch to share facial expression features with different contexts, thus promoting feature diversity to improve robustness. The proposed DSGCN framework achieves state-of-the-art results on the FERV39K, DFEW and AFEW benchmarks, and ablation experiments verify the effectiveness of each module.

Keywords: video-based facial expression recognition · graph convolutional networks · dynamic-static relation

1 Introduction

Currently, automatic recognition of facial expressions, which plays a crucial role in various human-computer interaction systems [1], including medical treatment, driver assistance, and other areas, has been a popular subject for researchers. The goal of facial expression recognition (FER) is to classify the input images into seven basic expressions: neutral, happy, sad, surprised, afraid, disgusted,

S. Rudinac et al. (Eds.): MMM 2024, LNCS 14555, pp. 42–55, 2024.
https://doi.org/10.1007/978-3-031-53308-2_4

and angry. According to different types of input data, the FER systems can be divided into image-based FER and video-based FER. Early FER methods mainly focused on image-based facial expressions. However, since facial expression is a dynamic process characterized by the interplay of muscle movements in different regions of the face, understanding the temporal sequence of expressions, plays a more important role than classifying static images. As a result, video-based FER research has received increasing attention in recent years.

According to different data scenarios, video-based FER datasets can be mainly divided into lab-controlled and in-the-wild. For the lab-controlled datasets, all video sequences are collected in a controlled laboratory environment, and the videos are relatively simple and free from occlusion. Representative datasets include CK+ [2], Oulu-CASIA [3], and MMI [4]. For the in-the-wild datasets (e.g., AFEW [5], Ferv39k [6], and DFEW [7]), video sequences are collected from real-world scenes, which are closer to natural facial events. Furthermore, the in-the-wild datasets are captured from thousands of subjects in complex scenes, which greatly increases the diversity of the data. Nowadays, the focus of video-based FER research has shifted from laboratory controls to challenges under field conditions.

Early methods for solving video-based FER were primarily based on handcrafted features, such as the LBP-TOP [8], STLMBP [9], and HOG-TOP [10]. In addition, Liu et al. [11] introduced a spatio-temporal manifold(STM) method to model the video clips, and Liu et al. [12] used different Riemannian kernels measuring the similarity distance between sequences. In recent years, deep learning based methods gradually replace traditional methods. Among these methods, the RNN-based method performed better at capturing the temporal relationship between frames, and the spatial self-attention emerged [13] was proposed as a powerful tool for guiding the extraction of image features and determining the importance of each local feature. However, these methods focused on limited attention features or relationships from a single perspective, thus neglecting large variations of different perspective expression patterns and data uncertainties. The 3D CNN-based method [14] was able to learn both spatial and temporal features in the sequence, but failed to effectively utilize long-distance attention-dependent information to extract rich emotional features. Meanwhile, CNN-based methods require stacking multiple layers of convolutional layers to enlarge the receptive field. However, this often leads to the loss of input information, increases computational load, and may even result in gradient vanishing issues.

Motivated by the above shortages of existing methods, in this paper we propose a novel dynamic-static graph convolutional network (DSGCN) for video-based FER. DSGCN consists of the Static-relational graph (SRG) and the Dynamic-relational graph (DRG). Specifically, in SRG, our method first focuses on the static spatial features extracted from the input facial expression images, and then constructs GCN for these features, the features from each frame are used as the vertex and the spatial similarity are used as the edges. Thus the constructed SRG strengthens the salient areas of the face through the dependencies of context nodes, and weakens the impact of in-the-wild factors (illumination changes, non-frontal head poses, facial occlusions) on the final recognition. In DRG our method first aggregates the features of the entire input video sequence to learn the dynamic temporal information, then constructs GCN on other video samples in the same batch. The sample nodes in the batch share features through similarity, improving the robustness of facial expressions extracted in a single situation, thus better dealing with complex and changeable real situations. At last, the video-based FER task is transferred into a node classification problem in the graphs constructed on the batches.

In summary, this paper has the following contributions:

(1) We propose DSGCN that simultaneously captures static spatial feature relationships, long-distance dynamic temporal dependencies and sample similarity relationships to gain efficient expression-related features.
(2) We present a graph-based approach for solving the task of video-based facial expression recognition by casting it as a node classification problem.
(3) Extensive experiments demonstrate DSGCN is able to outperform the baseline model significantly and achieve state-of-the-art results on three popular video-based FER datasets. Ablation studies verify the effectiveness of the composed modules (*i.e.*, SRG, DRG).

2 Related Work

2.1 Image-Based FER in the Wild

The Image-based FER mainly consists of three stages, namely face detection, feature extraction and expression recognition. In the face detection stage, methods such as MTCNN [15] and Dlib [16] are usually used to locate faces in complex situations. In the feature extraction stage, early methods mostly use hand-extracted features. Among them, texture-based features include HOG [10], Histograms of LBP [8], Gabor wavelet coefficients. At the same time, there are many methods of extracting features based on landmark points such as noses, eyes, and mouths, and using multiple feature combinations to obtain richer representations. Currently deep learning based methods are widely used. Fasel [17] found that shallow CNNs are robust to facial poses. Tang and Kahou et al. [18] used deep CNN for feature extraction and won the FER2013 and Emotiw2013 challenges respectively. Liu et al. [19] proposed a CNN architecture based on facial action units for expression recognition. The next stage after feature extraction is to feed the features into supervised classifiers such as support vector machines (SVM), softmax layers, and logistic regression to assign facial expression categories.

2.2 Video-Based FER in the Wild

In order to capture the spatio-temporal information in the video, methods based on CNN and RNN have emerged. Most of the CNN-RNN based DFER methods first use CNN to learn spatial facial features for each video frame, and then RNN processes the temporal information between video frames. Some methods use VGG or ResNet to extract spatial features, and long short-term memory (LSTM) or Gated Recurrent Unit (GRU) to extract temporal features. For example, Baddar et al. [20] proposed a pattern varied LSTM to encode spatio-temporal features that are robust to unseen changing patterns. For 3D CNN-based methods [14], spatial and temporal feature representations of video sequences are jointly extracted through 3D convolutions. Some 3D-CNN-based methods extract temporal and spatial features of video sequences through 3D convolutions. These methods [21] extract spatio-temporal facial features by directly adopting 3D-CNN, and such spatio-temporal features are usually combined with other types of facial features. Recently, Liu et al. [22] leveraged graph convolutional networks (GCNs) to learn frame-based features that focus on specific expression regions. Lee et al. [23] proposed a Multi-modal Recurrent Attention Network (MRAN) for learning spatio-temporal attention maps for robust DFER in the wild. Zhao et al. [24] first introduced the transformer to the DFER task, they designed CS-Former and T-Former for extracting spatial and temporal features.

2.3 GNN for Video Understanding

In recent years, transformer and GNN based methods have demonstrated excellent performance in the field of video understanding, especially in improving the performance of CNN/RNN-based methods. In the field of video understanding, GNNs have been applied in dialogue modeling, video retrieval, emotion recognition and action detection. There are also video representation frameworks that can be used for multiple downstream tasks. For example, Arnab et al. [25] created a fully connected graph using foreground nodes extracted from video frames in a sliding window fashion. They established connections between the foreground nodes and the context nodes of adjacent frames. Liu et al. [22] introduced the GCN layer in the general CNN-RNN based model of video-based FER, but they only focused on the relationship between frames, and did not focus on the similarity between samples. Differently, our work is dedicated to construct a graph structure that can capture more relationships.

3 Proposed Method

As shown in Fig. 1, the proposed DSGCN mainly consists of a static-relational graph (SRG) and dynamic-relational graph (DRG). The input of DSGCN are dynamically sampled fixed-length facial expression sequences from raw videos. SRG takes video series as input, dividing the video to single frames and extracting spatial facial features for each frame. Subsequently, SRG constructs a graph by using the spatial feature of each frame as nodes, the similarity between nodes as edges, thus capturing the long-distance dependencies of expressions. DRG aggregates the spatial feature sequence enhanced by SRG, and constructs GCN from other sample videos in the same batch, sharing feature information through the similarity between samples. Finally, the classification results are obtained by a full-connected (FC) layer.

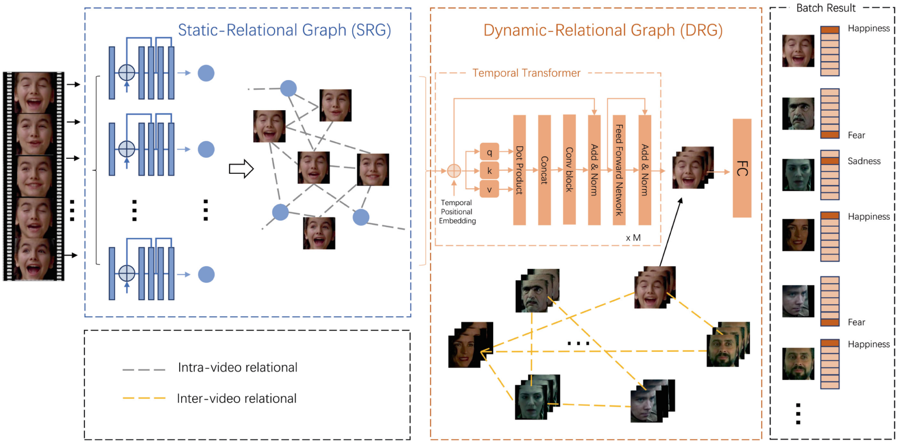

Fig. 1. The proposed model (DSGCN) architecture, which mainly consists of a Static-Relational graph (SRG) and a Dynamic-Relational graph (DRG).

3.1 Static-Relational Graph (SRG)

SRG mainly builds GCNs from frame nodes with rich spatial features. Given a facial expression video as input, a fixed-length sequence of facial expressions dynamically sampled from the raw video sequence is fed into the model. The frames in the sequence are first transformed to features carrying rich facial spatial information through the spatial network module, and then GCNs are constructed based on the features to strengthen the salient facial expression regions.

Static Spatial Feature: Fixed-length clip $X \in \mathbb{R}^{T\times 3\times H\times W}$ are obtained as input by dynamically sampling raw video. Specifically, we split the video sequence into S segments, and randomly select V frames in each segment. We thus obtain an input clip of fixed length $T = S \times V$.

For building static-relational graph, extracting rich spatial representation from the frame, we use a Spatial Transformer [24]. The Spatial Transformer consists of five convolution blocks and N spatial encoders. The previous four convolution blocks, including conv1, conv2, conv3 and conv4, are used to extract local facial spatial features $M \in \mathbb{R}^{C\times H'\times W'}$. After this, we flatten the feature and add positional information $P_{spatial}$ to feed it into N spatial encoder. The spatial encoders consist of a multi-head self-attention and feed forward network to model global spatial relationships. The final convolution block conv5 is used to refine the final facial features. Therefore, input the Spatial Transformer of the t-length clip, and the output is $F \in \mathbb{R}^{T\times f}$ that carries sufficient spatial information.

Intra-Video Graph: In order to capture the long-range dependencies of facial regions in videos, we propose a graph-based module to capture expression changes. We construct a GCN layer by obtaining T features with spatial-temporal relations from the previous module, and model the contextual correlation by learning the dynamic adjacency matrix A. All nodes tend to be influenced by expression informative frames and update themselves as more contributing ones.

The inputs are representation maps $\hat{F} = \{\hat{f}_1, \ldots, \hat{f}_T\}$extracted by spatial transformer from the original video. To begin with, we use cosine similarity coefficient to calculated the similarity between different representations as:

$$cossim(f_i, f_j) = \frac{f_i * f_j}{\|f_i\| \|f_j\|} \tag{1}$$

At the same time, we construct the adjacency matrix A through the cosine similarity coefficient, and $A_{i,j}$ represents the similarity between node i and node j. And in each time step, as the node features are updated, the adjacency matrix A will also update the similar state between nodes.

$$A_{i,j} = cossim(f_i, f_j) \tag{2}$$

$i, j \in \{1, 2, \ldots, T\}$. then, we employ GCN as:

$$F^{l+1} = \bar{D}^{-\frac{1}{2}} \bar{A} \bar{D}^{-\frac{1}{2}} F^l W^l \tag{3}$$

where l represents the lth time step, $\bar{A} = A + I$ is the sum of un-directed graph A and the identity matrix, $\bar{D}$ is the diagonal matrix from A, which is $\bar{D}_{i,i} = \sum_j A_{i,j}$. F^l and F^{l+1} are the corresponding input and output representations on the $l_t h$ level, and W^l are the trainable parameters on this level. At each time step, the GCN layer shares the features of each node to neighbor nodes based on the adjacency matrix A, and accepts update messages from neighbor nodes.

3.2 Dynamic-Relational Graph (DRG)

Dynamic Temporal Feature: Using the output F' from the SRG as input, DRG aims at capturing the dynamic temporal relation for the feature nodes that have obtained spatial information, and mining the facial expression movement information between nodes. In the method, we first use the implemented Temporal Transformer [24]. The Temporal Transformer consists of M temporal encoders, each of which includes a multi-head self-attention and a feed-forward network. For T spatial features X' from the Spatial relation graph, they will be input to the temporal encoder after adding position information $P_{temporal}$. Through the multi-head self-attention and a feed-forward network in the temporal encoder, the global temporal information is modeled to output features h with rich spatial-temporal information.

Inter-Video Graph: Not limited to learning contextual relations in videos, we then extend the DRG module to learn the similarity between the input video samples. Our module accepts B video samples in the same batch, and the features of each video are transformed into a feature h carrying rich spatio-temporal information through the above steps, then we construct GCNs on these features. Our graph structure learns different scene knowledge of similar expressions for each video node in a single scene by sharing video samples of different expressions.

The inputs are representation maps $\hat{H} = \{\hat{h_1}, \ldots, \hat{h_B}\}$extracted by temporal transformer from the sample video under the same batch. We will construct the adjacency matrix A based on Eqn. (2), and update the video nodes in the same batch based on Eqn. (3). After l rounds of updating node features, each node successfully learns different scene knowledge of similar expressions to deal with the large variations of expression patterns and data uncertainties.

Node Classification: In the previous steps, we have described our graph construction procedure that converts a batch video into a graph where each node has its own spatio-temporal feature vector. During the training process, we feed all videos in a batch simultaneously into the proposed model, and add fc layers at the end of the outputs, transferring the original video-based facial expression recognition into a seven-category node classification problem in the constructed graph.

4 Experiments

4.1 Datasets

FERV39k. [6] Currently represents the largest in-the-wild DFER dataset, containing 38,935 video clips collected from four different scenarios, which can be recursively divided into 22 fine-grained scenes, such as daily life, talk shows, business, and crime. All scene video clips are randomly shuffled and 80% of clips are allocated to the training set, while 20% of clips are reserved for the test set to avoid dataset overlaps. Therefore, in order to conduct a fair comparison, we directly use the training and testing sets divided by FERV39k.

DFEW [7] is a database that contains 16,372 video clips from more than 1,500 movies. All samples have been divided into five equally sized parts (fd1 fd5). Five-fold cross-validation is used as the evaluation scheme. In each fold, a portion of the samples is used for testing while the remaining data is reserved for training. Finally, all predicted labels are used to compute an evaluation metric by comparing them with the ground truth.

AFEW. [5] Dataset serves as the evaluation platform for the annual EmotiW challenge from 2013 to 2019. AFEW contains 1809 video clips collected from different movies and TV series. Consistent with DFEW, each video clip in AFEW is assigned to one of seven basic expressions. The test clips are not publicly available, so we train our model on train clips and test on validation clips.

Table 1. Comparison with state-of-the-art methods on FERV39k.

Methods	Accuracy of Each Emotion (%)							Metrics (%)	
	Happiness	Sadness	Neutral	Anger	Surprise	Disgust	Fear	UAR	WAR
C3D	48.20	35.53	52.71	13.72	3.45	4.93	0.23	22.68	31.69
P3D	61.85	42.21	49.80	42.57	10.50	0.86	5.57	30.48	40.81
R2Plus1D	59.33	42.43	50.82	42.57	16.30	4.50	4.87	31.55	41.28
3DR18	57.64	28.21	59.60	33.29	4.70	0.21	3.02	26.67	37.57
R18+LSTM	61.91	31.95	61.70	45.93	14.62	0.00	0.70	30.92	42.59
VGG13+LSTM	66.26	51.26	53.22	37.93	13.64	0.43	4.18	32.42	43.37
Two C3D [6]	54.85	52.91	60.67	31.34	5.96	2.36	6.96	30.72	41.77
Two R18+LSTM [6]	59.00	45.87	61.90	40.15	9.87	1.71	0.46	31.28	43.2
Two VGG13+LSTM [6]	69.65	47.31	52.55	47.88	7.68	1.93	2.55	32.79	44.54
Former-DFER [24]	65.65	51.33	56.74	43.64	21.94	8.57	12.52	37.20	46.85
STT [26]	69.77	47.81	59.14	47.41	20.22	**10.49**	9.51	37.76	48.11
NR-DFERNet [27]	69.18	54.77	51.12	49.70	13.17	0.00	0.23	33.99	45.97
DSGCN	**86.90**	**61.95**	**72.32**	**55.68**	**31.19**	9.21	**16.24**	**47.64**	**59.88**

4.2 Implementation Details

Training Setting: For all the three datasets, we train our model from scratch with a batch size of 32, initialize the learning rate to 0.01, and divide it by 5 every 50 epochs. Due to the small number of data samples in AFEW dataset, in order to make a fair comparison, we first pre-train our model and other models on DFEW (fd1), and then fine-tune on AFEW with the same setting.

Evaluation Metrics: Without loss of generality, We choose Unweighted Average Recall (UAR, i.e. the accuracy of each category divided by the number of

Table 2. Comparison with state-of-the-art methods on DFEW.

Methods	Metrics(%)	
	UAR	WAR
3DR18	46.52	58.27
R18+LSTM	51.32	63.85
R18+GRU	51.68	64.02
EC-STFL [7]	45.35	56.51
Former-DFER [24]	53.69	65.70
EST [28]	53.43	65.85
STT [26]	54.58	66.65
NR-DFERNet [27]	54.21	68.19
GCA+IAL [29]	55.71	69.24
DPCNet [30]	**57.11**	69.24
DSGCN	57.06	**70.57**

Table 3. Comparison with state-of-the-art methods on AFEW.

Methods	Metrics(%)	
	UAR	WAR
C3D	43.75	46.72
I3D-RGB	41.86	45.41
R(2+1)D	42.89	46.19
3DR18	42.14	45.67
R18+LSTM	43.96	48.82
Former-DFER [24]	47.42	50.92
EST [28]	49.57	54.26
STT [26]	49.11	54.23
NR-DFERNet [27]	48.37	53.54
DPCNet [30]	47.86	51.67
DSGCN	**60.46**	**65.49**

categories, regardless of the instances of each category) and weighted average recall (WAR, i.e. accuracy) as the metrics.

4.3 Comparison with State-of-the-Arts

In this section, we compare our best results with current state-of-the-art methods on the FERV39k, DFEW and AFEW benchmarks.

As shown in the Table 1, we compare our method with other state-of-the-art methods on the FERV39k dataset, including C3D, P3D, R2Plus1, 3DR18, R18+LSTM, VGG13+LSTM, Two C3D [6], Two R18+LSTM [6], Two VGG13+LSTM [6], Former-DFER [24], STT [26], NR-DFERNet [27]. DSGCN improvements of 9.88% and 11.77% in UAR and WAR than the previous state-of-the-art method STT. Moreover, we show the performances on each expression in the Table. As can be seen, our method achieve the best results on most of the expressions, only slightly lower than STT on Disgust with a gap of 1.28%. At the same time, in Table 1, we can see that most of the methods perform poorly in "disgust" and "fear", which we believe is caused by insufficient training data in the original datasets.

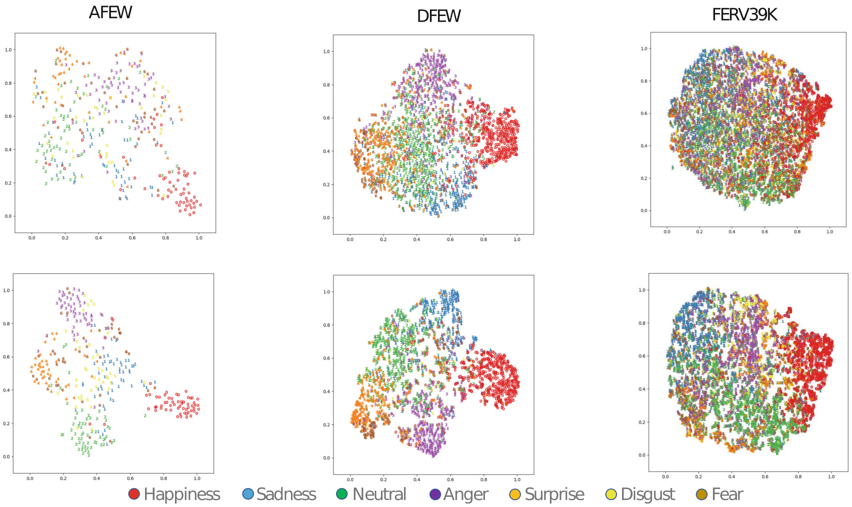

Fig. 2. Illustration of feature distribution learned by the Former-DFER [24] (top) and DSGCN (bottom) on three datasets.

For the DFEW data set, the experiment compared 3DR18, R18+LSTM, R18+GRU, EC-STFL [7], Former-DFER [24], EST [28], STT [26], NR-DFERNet [27], GCA+IAL [29], and DPCNet [30] under 5-fold cross-validation. As shown in Table 2, DSGCN outperforms the comparison methods on the WAR metric, and is very close to the current state-of-the-art method DPCNet on the UAR metric.

Specifically, we have a 1.33% improvement on WAR and only a 0.05% reduction on UAR compared to DPCNet. It should be noticed that DFEW also has a imbalanced data distribution. The proportions of "disgust" and "fear" sequences are 1.22% and 8.14%, which is the reason why our method achieve a relatively low performance in UAR.

For the AFEW dataset, all methods are first pre-trained on DFEW (fd1) and then fine-tuned on AFEW with the same settings. Our method compares C3D, I3D-RGB, R(2+1)D, 3DR18, R18+LSTM, Former-DFER [24], EST [28], STT [26], NR-DFERNet [27], DPCNet [30]. The comparative performance shown in Table 3 shows that DSGCN achieves the best results on both UAR and WAR. In particular, our method shows an improvement of 10.89% and 11.23% in UAR and WAR than the previous state-of-the-art method EST.

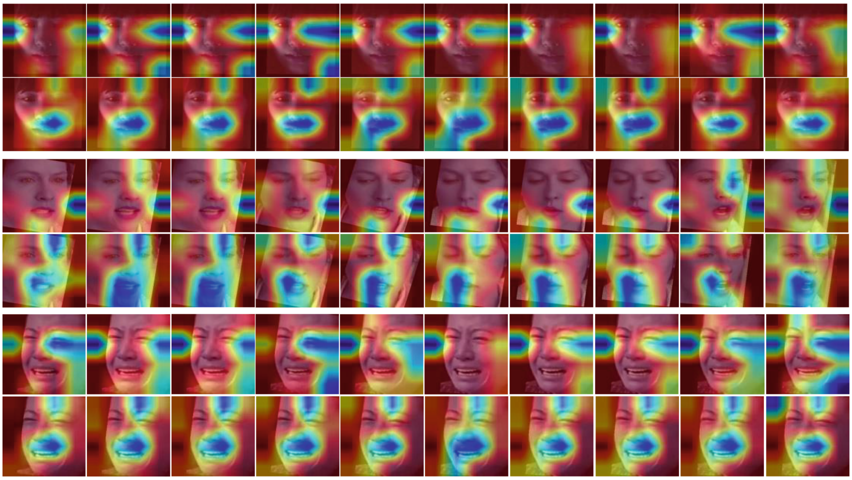

Fig. 3. Visualization of the learned feature maps. There are three sequences are presented, which including the facial expression of Neutral, Anger and Sadness, respectively. For each sequence, the images in the first row are heat-maps generated by the Former-DFER, and the images in the second row are heat-maps generated by DSGCN.

4.4 Visualization Results

We utilize t-SNE [31] to analyze the feature distribution learn by the Former-DFER and DSGCN on three datasets. As shown in Fig. 2, it is obvious that the feature distribution of each category learned by our method is tighter, and the boundaries between different categories are more obvious. This shows that our method can better discriminate different facial expressions in feature level. Furthermore, we conduct experiments to visualize the learned facial feature maps, as shown in Fig. 3, we used neutral, angry, and sad three types of expressions to compare with Fomer-DFER. For the first neutral expression sequence, although there is no significant expression behavior, our method still pays more attention to facial regions. In the second angry expression sequence and the third sad expression sequence, our method pays more attention to the facial regions such as mouth and eyes that contain more emotional information. In the second sequence where the subject has large head pose changes, our method always locks on the subject's face region compared to the comparison method.

Table 4. Ablation study to evaluate the effectiveness of different modules in our proposed method.

Methods	FERV39K(%)		DFEW(%)		AFEW(%)	
	UAR	WAR	UAR	WAR	UAR	WAR
Baseline	37.20	46.85	53.69	65.70	47.42	50.92
SRG	44.64	57.47	55.29	67.56	58.95	64.15
DRG	44.35	57.74	56.43	68.38	59.90	64.69
DSGCN	**47.64**	**59.88**	**57.06**	**70.57**	**60.46**	**65.49**

4.5 Ablation Study

We conduct extensive ablation studies on the three video-based FER datasets to demonstrate the effectiveness of different components of our proposed method. Including the individual part of SRG and DRG, as well as the final DSGCN. The Former-DFER is employed as the baseline. As shown in Table 4, our STRG achieves the WAR and UAR of 40.43%/54.81%, 55.56%/67.04%, 57.67%/62.80% on three datasets, which outperforms some existing methods because of the spatio-temporal features we learned. in one hand, by using SRG we have achieved obvious improvements in WAR and UAR compared to the baseline. This proves that SRG can effectively enhance facial expressions by learning the similarity of expressions at different moments in the same video, and provide more robust features for subsequent extraction of temporal information. In other hand, through the propagation and enhancement of spatio-temporal features, DRG outperforms the baseline to varying degrees on the three datasets. The most significant improvements are in AFEW, where CRG exceeds baseline by 12.48% and 13.77% on WAR and UAR. This proves that DRG can indeed capture the correlation between different sample expressions to strengthen the current expression. We notice solely using DRG performs slightly better than using SRG, the reason is dynamic features from other video sequences can better improve the robustness of node features than in the same video sequence. Finally, in the complete method DSGCN, all indicators in the three datasets reach the highest, the results prove our method can indeed learn both the spatial and temporal relationship of the input video facial expressions.

5 Conclusion

This paper proposes a novel dynamic-static graph convolutional network (DSGCN) for dynamic facial expression recognition in-the-wild scenarios. Specifically, the proposed DSGCN mainly consists the static-relational graph (SRG) and the dynamic-relational graph (DRG), it can capture multi-level relationships among the input video sequences, including spatial relationship, temporal relationship, context relationship and sample relationship. Extensive experiments with previous methods show that the proposed DSGCN achieves state-of-the-art results on three popular dynamic FER benchmarks. The abundant ablation studies have validated the effectiveness of each part in DSGCN. Moreover, the visualization results of facial features demonstrate that DSGCN can pay more attention to the salient facial regions. The visualization results of the feature distribution show that the method can better discriminate the learned face features. Furthermore, comparisons with previous methods show that DSGCN achieves state-of-the-art results on three popular dynamic FER benchmarks.

In future work, based on our DSGCN framework, we will further expand it to Micro-Expression Recognition, Pose Prediction, Person Recognition and other fields. Additionally, we plan to integrate our DSGCN framework with self-supervised learning, encouraging the model to learn potential internal relationships in a large amount of unlabeled data, thereby alleviating the impact of imbalances in facial data.

Acknowledgements. This work was supported in part by Zhejiang Provincial Natural Science Foundation of China (No.LDT23F0202, No. LDT23F02021F02, No. LQ22F020013) and the National Natural Science Foundation of China (No. 62036009, No. 62106226).

References

1. Plass-Oude Bos, D., Poel, M., Nijholt, A.: A study in user-centered design and evaluation of mental tasks for BCI. In: Lee, K.-T., Tsai, W.-H., Liao, H.-Y.M., Chen, T., Hsieh, J.-W., Tseng, C.-C. (eds.) MMM 2011. LNCS, vol. 6524, pp. 122–134. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-17829-0_12
2. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-workshops. pp. 94–101. IEEE (2010)
3. Zhao, G., Huang, X., Taini, M., Li, S.Z., PietikäInen, M.: Facial expression recognition from near-infrared videos. Image Vis. Comput. **29**(9), 607–619 (2011)
4. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: 2005 IEEE International Conference on Multimedia and Expo, P. 5. IEEE (2005)
5. Dhall, A., Goecke, R., Lucey, S., Gedeon, T., et al.: Collecting large, richly annotated facial-expression databases from movies. IEEE Multimedia **19**(3), 34 (2012)
6. Wang, Y.: 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 (2022)

7. Jiang, X.: 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 (2020)
8. Wang, L., He, X., Du, R., Jia, W., Wu, Q., Yeh, W.: Facial expression recognition on hexagonal structure using LBP-based histogram variances. In: Lee, K.-T., Tsai, W.-H., Liao, H.-Y.M., Chen, T., Hsieh, J.-W., Tseng, C.-C. (eds.) MMM 2011. LNCS, vol. 6524, pp. 35–45. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-17829-0_4
9. Huang, X., He, Q., Hong, X., Zhao, G., Pietikainen, M.: 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 (2014)
10. Chen, J., Chen, Z., Chi, Z., Fu, H.: 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 (2014)
11. Liu, M., Shan, S., Wang, R., Chen, X.: Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1749–1756 (2014)
12. Liu, M., Wang, R., Li, S., Shan, S., Huang, Z., Chen, X.: Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In: Proceedings of the 16th International Conference on Multimodal Interaction, pp. 494–501 (2014)
13. Aminbeidokhti, M., Pedersoli, M., Cardinal, P., Granger, E.: Emotion recognition with spatial attention and temporal softmax pooling. In: Karray, F., Campilho, A., Yu, A. (eds.) ICIAR 2019. LNCS, vol. 11662, pp. 323–331. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-27202-9_29
14. Fan, Y., Lu, X., Li, D., Liu, Y.: Video-based emotion recognition using cnn-rnn and c3d hybrid networks,. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 445–450 (2016)
15. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)
16. Amos, B., Ludwiczuk, B., Satyanarayanan, M., et al.: Openface: a general-purpose face recognition library with mobile applications. CMU School Comput. Sci. **6**(2), 20 (2016)
17. Fasel, B.: Robust face analysis using convolutional neural networks. In: 2002 International Conference on Pattern Recognition, vol. 2, pp. 40–43. IEEE (2002)
18. Kahou, S.E., et al.: 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 (2013)
19. Liu, M., Li, S., Shan, S., Chen, X.: Au-inspired deep networks for facial expression feature learning. Neurocomputing **159**, 126–136 (2015)
20. Baddar, W.J., Ro, Y.M.: 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 (2019)
21. Lee, J., Kim, S., Kim, S., Park, J., Sohn, K.: Context-aware emotion recognition networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10143–10152 (2019)
22. Liu, D., Zhang, H., Zhou, P.: Video-based facial expression recognition using graph convolutional networks. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 607–614. IEEE (2021)

23. Lee, J., Kim, S., Kim, S., Sohn, K.: Multi-modal recurrent attention networks for facial expression recognition. IEEE Trans. Image Process. **29**, 6977–6991 (2020)
24. Zhao, Z., Liu, Q.: Former-dfer: dynamic facial expression recognition transformer,. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 1553–1561 (2021)
25. Arnab, A., Sun, C., Schmid, C.: Unified graph structured models for video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8117–8126 (2021)
26. Ma, F., Sun, B., Li, S.: Spatio-temporal transformer for dynamic facial expression recognition in the wild, arXiv preprint arXiv:2205.04749 (2022)
27. Li, H., Sui, M., Zhu, Z., et al.: Nr-dfernet: noise-robust network for dynamic facial expression recognition, arXiv preprint arXiv:2206.04975 (2022)
28. Liu, Y., Wang, W., Feng, C., Zhang, H., Chen, Z., Zhan, Y.: Expression snippet transformer for robust video-based facial expression recognition. Pattern Recogn. **138**, 109368 (2023)
29. Li, H., Niu, H., Zhu, Z., Zhao, F.: Intensity-aware loss for dynamic facial expression recognition in the wild, arXiv preprint arXiv:2208.10335 (2022)
30. Wang, Y., et al.: 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 (2022)
31. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. J. Mach. Learn. Res. **9**(11) (2008)