MTFNet: Multi-Scale Transformer Framework for Robust Emotion Monitoring in Group Learning Settings

Yi Zhang[1†], FangYuan Liu[1†], JiaJia Song[1], Qi Zeng[2] and Hui He[1*]
1 Beijing Normal University, Zhuhai, ZhuHai, China
2 Beijing Normal University, Beijing, China

E-mail: hhdpc@outlook.com Tel: +86-17328420173

***Abstract*—Identifying students' learning states in authentic classroom settings is a prominent topic in educational technology. This study addresses the challenges posed by complex facial environments and the scarcity of data in such settings. we propose a Multi-Scale Transformer with Frame Shuffled Order Predict Network (MTFNet), based on a spatial-temporal feature extraction structure, to perform effective learning-related facial expression recognition in a primary school classroom. Specifically, we combine a Multi-Scale Facial Feature Fusion Module (MFFF) based on Grouped Spatial Convolution(GS Conv) to effectively capture multi-level facial features and improve the model's robustness in complex environments. Additionally, a Frame-wise Shuffle Order Prediction Module (FSOP) is introduced to enhance the model's ability to understand the dynamic changes of emotional intensity by predicting the emotion expression sequence. Experiments on both the DFEW dataset and our dataset demonstrate excellent performance and generalization in real-world applications.**

I. INTRODUCTION

Facial expressions are the most powerful and direct means for humans to convey emotions [1]. In recent years, facial expression recognition (FER) has garnered significant research interest due to its potential applications in marketing [2], education [3], and safe driving [4]. For instance, students' facial expressions can reflect their feelings about the course content and indicate their engagement in group discussion, which allows teachers to verify whether students understand the material effectively and promptly, enabling timely adjustments to the course schedule [5][6][7]. However, unlike the laboratory settings, there are lots of challenges in detecting and recognizing expressions in the wild, such as occlusions and low resolution, limiting the application of artificial intelligence in the real world. Therefore, some studies focused on adapting models' structure or enhancing data quality to improve the inference efficiency and accuracy on wild face datasets [8][9][10].

From the perspective of the model, it is worth noting that the presentation of facial expressions is often a dynamic process, so relying solely on a static single image as input will confuse the model on similar expressions. Compared with traditional hand-crafted feature extractions such as Histogram of Oriented Gradients (HOG) [11] and Local Binary Pattern (LBP) [12], deep learning methods have got more excellent grades in a large number of experiments. Specifically, CNN-based models such as ResNet [13] can easily extract spatial features from each frame while RNN-based models such as LSTM [14] have the ability to find temporal dependencies between frames. Combining the two methods is an excellent way to effectively model the dynamic facial expression process. However, a more effective approach today involves using transformer-based models, which consistently achieve state-of-the-art performance due to their robust global information perception capability [8][15][16]. These models can capture crucial information in the input without depending on sequence order. Therefore, it is possible to train the transformer-based model by shuffling the order of frames.

From the perspective of the dataset, while previous studies have demonstrated the potential benefits of deep learning models in complex situations [17][18][19], there is still insufficient evidence to support their effectiveness in classroom contexts. In fact, the large-scale classrooms in China create an urgent demand for a model that can track students' learning-related emotions during class, assisting teachers in better monitoring each student's emotional engagement. Although in the field of education, the types of emotions generated by students when they are learning are very complicated (e.g. bored, pride, and confused) [20], most of the literature on affective computing in education is still based on Ekman's discrete emotion theory [21], which is a general and convenient way to categorize emotions but does not always provide valuable and practical insights for users. Therefore, it is essential to develop a learning emotion dataset aligned with learning engagement theory, grounded in both theoretical and empirical research in the field of education.

†: Equal contribution.
*: Corresponding author.

All in all, the main contributions of this paper can be summarized as follows:

- Proposing a MFFF module based on GS convolution to effectively capture features from video clips with occlusions and uneven lighting, which improves the performance of the model in practical applications.
- Adapting a FSOP module to guide the model to focus on changes between frames, thereby enhancing the transformer's ability to model effect on dynamic changes in facial expressions.
- Providing a small-scale video dataset, which records student's performance in a primary school classroom and includes four discrete learning-related emotions: interesting, happy, bored and confused.
- Identifying that proposed MTFNet achieves state-of-the-art performance on the evaluated datasets (i.e., DFEW and our datasets). We will release our source code upon acceptance.

II. Related Work

In this section, we review some related works of our proposed method for facial expression recognition in educational scenarios.

A. Emotion recognition based on facial expressions

In the 1970s, Ekman and Friesen found that the way humans perceive certain emotions is the same, regardless of cultural and racial differences, and defined six basic emotions: happiness, anger, disgust, fear, sadness and surprise [21]. Subsequently, in 1978, Ekman and colleagues further refined their researches on facial expressions and proposed the Facial Action Coding System (FACS) based on modular modeling of facial expression recognition muscles [22]. This system has a broader and more diverse range of expression portrayals, which has greatly promoted the research progress of facial expression recognition. Currently, facial expression recognition is mainly divided into two categories: single-frame expression recognition (SER) and dynamic expression recognition (DER).

SER is to recognize the expression state of a single-frame image with significant facial changes. In the early days, due to the limitations of the small scale of data and the poor calculated performance of computers, researchers preferred to study the expression recognition by single frame, and achieved good research results on various data sets such as CK+ and Affectnet [23][24][25].

DER is to recognize the change process of a complete facial expression (e.g. neutral -> start -> apex -> end -> neutral). The process is shown in the Fig. 1 . In 1978, Suwa and colleagues began to recognize facial expressions from facial video in order to make the recognition results more representative, thus opening up the field of DER. Early researchers used manual operators (Gabor, LBP, Haar, etc.) to extract specific features for recognition [26][27][28]. In 2011, Gedeon and colleagues proposed the AFEW dataset, which contains 1,743 video clips from film and television works. This greatly promoted the application of deep learning in DER [29]. A large number of researchers applied deep learning methods to the extraction of facial expression features such as CNN-based methods and RNN-based methods [30][31][32]. In 2017, the Google team used a single attention mechanism to get great success in machine translation tasks [15], which then was applied to spatial and temporal feature modeling. For instance, Zengqun and colleagues designed a spatial and a temporal self-attention module at the level of model structure to fully model the expression features from the video [23] . In addition, some researchers have also combined temporal features with spatial features for processing, such as C3D and MAE-DEFR, which also have achieved great results on multiple dataset [33], [34].

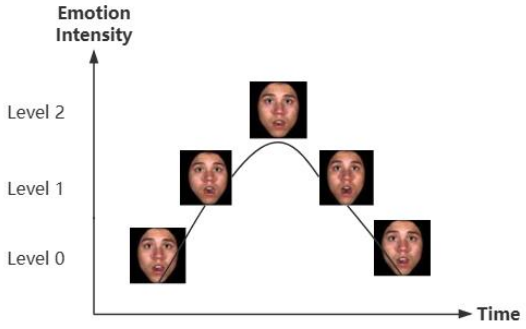

Fig. 1 Dynamic evolution of facial expressions chart

B. Identification of students' learning in classroom contexts

Currently, the recording and evaluation of students' learning status mainly adopts self-evaluation and classroom reports, which refers to students' evaluation of their own learning status after the course ends. With the rise of research on physiological signals in recent years, biofeedback technology has been used to obtain physiological indicators to analyze the learning status of learners. For instance, Wang Pengli and colleagues monitored learners' EEG signals to monitor their attention in class through brain-computer interface devices [35]. Moreover, the development of camera equipment and computer vision technology in recent years allows researchers focusing on identifying students' emotions through images, that is, identifying the emotional category based on changes in the learner's face, gestures and eye movements over a period of time. For instance, Sun Bo and colleagues [36] moved the learning scenario from online to offline and proposed a teacher-student emotional interaction system, which realized the emotion recognition and intervention functions of learners. Zhan Zehui et al. tracked learners' eye movement status and integrated it with emotional characteristics to build an intelligent agent learner emotion and cognitive model [37]. In summary, these systems enable teachers to monitor students' learning status in real time, allowing for timely intervention when issues arise.

III. Methodology

In this section, we introduce the proposed framework in detail, including three modules: Multi-Scale Facial Feature Fusion module (MFFF), Temporal-Encoder with FSOP and Design of a Joint Loss Function.

A. Overview of the Proposed Framework

An overview of the proposed framework is presented in Fig. 2, which consists of three modules. First, For the input video V, the frame sequence $\{F_1, \cdots, F_n\}$ is extracted and then subsequently shuffled into $\{F_m, \cdots, F_t\}$. This shuffled sequence is fed into a MFFF module to extract frame-wise facial features

$\{f_m, \cdots, f_t\}$. Following this, temporal feature extraction is performed by adding temporal positional encodings $\{em_m, \cdots, em_t\}$ to derive the video features. Finally, these features are used for emotion classification module and frame order prediction module.

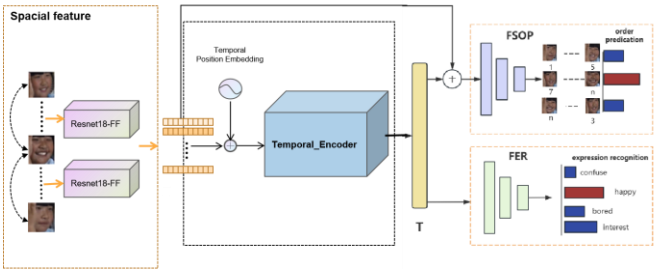

Fig. 2 Overview of the proposed framework for MTFNet from videos.

B. *Multi-Scale Facial Feature Fusion Module*

In real-world environments, students' facial emotion is often affected by lighting variations, occlusions, and poses changes. When the prominent features of a subject's face are obstructed, relying solely on deep, high-semantic features can severely impact the model's recognition performance. To mitigate the recognition challenges posed by occlusion, we employed a simple and efficient feature fusion strategy. Unlike the commonly used feature pyramid methods in object detection tasks, this approach focuses more on integrating features at different semantic levels rather than on positional information transmission. The model structure is shown in the Fig. 3. Given ResNet18's excellent generalization capabilities, it is used as the backbone network. The model includes two sets of feature fusion modules. The first set integrates the low-semantic features from stages 1 to 3 of ResNet18, while the second set merges the fused features of the first set with the high-semantic features from stages 3 to 4, ultimately obtaining multi-level facial features. As shown in the Fig. 4, each fusion unit comprises four functional channels, each utilizing two GS convolution modules. This module combines the original convolution kernels with depthwise separable convolutions, ensuring feature extraction capability while reducing computational complexity. The GS convolution [39] structure is shown in the Fig. 5.

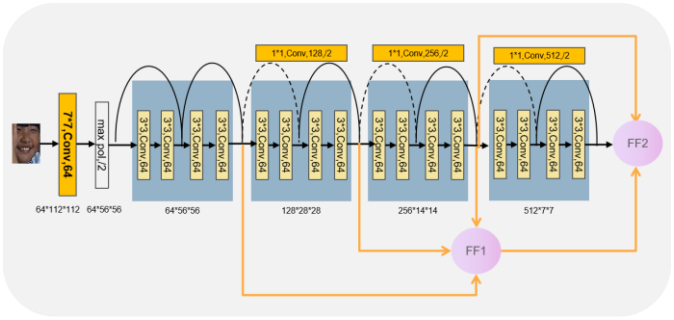

Fig. 3 ResNet18-FF Model Architecture Diagram

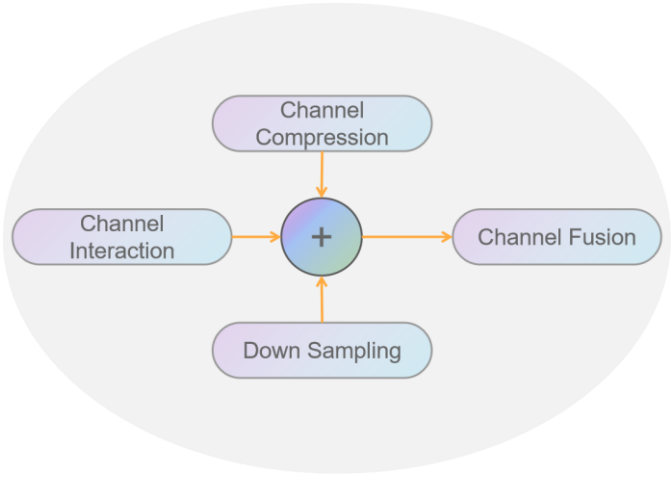

Fig. 4 Feature Fusion Module Structure Diagram

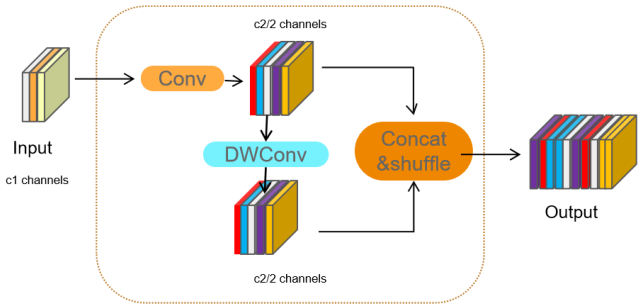

Fig. 5 GS Convolution Diagram

C. *Temporal-Encoder with FSOP*

In this module, leveraging the global modeling capabilities of the self-attention mechanism, we introduced the Temporal-Encoder for dynamic modeling of emotion sequences. The model structure is shown in the Fig. 6. Based on the random frame features $\{f_m, \cdots, f_t\}$ extracted by the Spatial Feature module, we added temporal positional encoding and performed multi-head attention for dynamic emotion feature modeling to obtain emotion classification features of the frame segments. However, in practical applications, the multi-head attention mechanism is easily influenced by the intensity variations in facial features, leading to neglecting the overall motion features. For instance, when peak frames exhibit strong noise interference, the extraction of dynamic features is severely affected. To address this issue, Liu et al. proposed the Frame-wise Shuffle Order Prediction module (SSOP), which divides emotion video sequences into multiple frame segments and performs order prediction at the segment level. While this method alleviates the problem, it is computationally too complex for real-time emotion prediction tasks. In this paper, we propose a Frame-wise Shuffle Order Prediction (FSOP) module, which performs order prediction at the frame level.

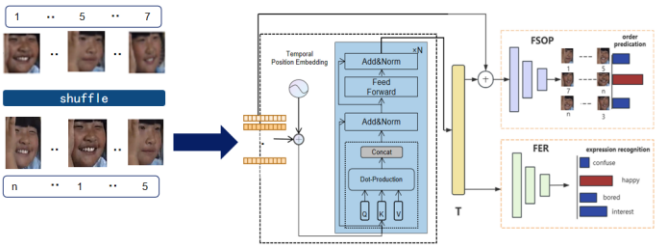

Fig. 6 Temporal-Encoder with FSOP

D. *Learning Criteria*

To supervise the learning of the proposed framework for expression recognition, we firstly consider cross entropy loss for Facial Expression Classification Module and Prediction of Frame Sequence Order, respectively. The cross entropy loss is defined as:

$$L_{cls} = -\sum_{i=1}^{N} y_i * \log(\hat{y}_i) + (1 - y_i) * \log(1 - \hat{y}_i) \quad (1)$$

$$L_{fsp} = -\sum_{i=1}^{N} S_i * \log(\hat{S}_i) + (1 - S_i) * \log(1 - \hat{S}_i) \quad (2)$$

Where N denotes the number of samples, yi denotes the true label, $\hat{y}_i$ denotes the predicted class probability, S_i denotes the true ordinal category, $\hat{S}_i$ denotes the predicted ordinal category probability.

Therefore, the overall loss function is defined as follows:

$$Loss = L_{cls} + aL_{fsp} \quad (3)$$

Where α denotes the weight of the proportion of category prediction results, Based on experimental results, the value of a was chosen to be 0.1428.

IV. Experiment

In this section, we present the experiments for evaluating the proposed framework for facial expression recognition. Specifically, i) we introduce the adopted databases in our experiments; ii) we demonstrate the evaluation indexes to assess models' performance in datasets; iii) we present the implementation details of the proposed framework; iv) we show and analyze the comparisons of our method with other state-of-the-art methods, as well as the ablation studies for investigating the effects from the different designs in our proposed framework.

A. Datasets

In our experiments, we adopted two datasets to evaluate the performance of the proposed framework, including AFEW and our dataset. Examples of the expressive faces from these databases are presented in Fig. 7.

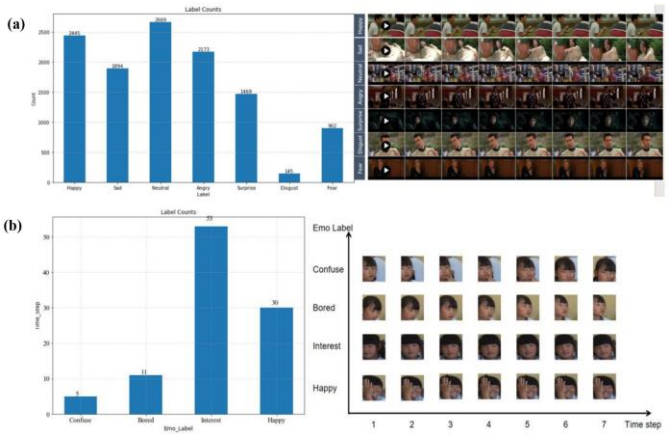

Fig. 7 Examples from (a) DFEW and (b) our dataset

The DFEW database consists of 16,372 video clips extracted from over 1,500 different movies. It is a large-scale, unconstrained dynamic facial expression database with a total of 12,059 single-label video clips. The dataset uses seven basic labels: Fear, Disgust, Surprise, Angry, Neutral, Sad, and Happy. Notably, the dataset encompasses various extreme scenarios such as lighting variations, occlusions, and pose changes, highlighting the significant challenges associated with accurate facial expression recognition in such complex environments.

Our dataset consists of emotion recognition data collected specifically for group learning environments, addressing a gap in existing datasets both domestically and internationally. In collaboration with educational experts, we defined four common and analytically useful emotional categories: interested, happy, bored, and confused, following extensive literature review, field observations, and expert discussions. Subsequently, we recorded five sets of 10 classroom group discussion videos, each approximately 7 minutes long, in primary and secondary schools. (This program was approved by the Ethics Committee of the authors' university (No. BNU202201100014). Informed consent was obtained from all individual participants included in the study and their parents.) The videos were processed into image frame sequences, and single-frame images with distinct emotional features were labeled by experts. Finally, a program was used to interpolate and extract seven frames spanning 2 seconds before and after the labeled frame to form video clips. After obtaining a large number of video frames, further screening was conducted to ensure dataset quality, such as removing unclear or blurred facial images and retaining those with complex background information. Fig. 7 shows examples of the retained video frames.

B. *Evaluation*

To assess the ability of proposed framework and other state-of-the-art methods, we employ three evaluation indexes: confusion matrix, weighted accuracy rate, and unweighted accuracy rate.

- **Confusion matrix**

The confusion matrix is often used for evaluating multi-class image classification. Each column represents a predicted class, with the total number in each column indicating the number of

instances predicted for that class. Each row represents the actual class, with the total number in each row indicating the number of instances that actually belong to that class.

- **Weighted accuracy rate**

Weighted Accuracy Rate (WAR) takes into account the weight of each class by multiplying the accuracy of each class by the number of samples in that class to calculate the weight. This approach ensures that both major and minor classes contribute appropriately when evaluating the overall performance of the model. Given the severe imbalance in the different expression categories in the two datasets, this metric is more representative for this task. The calculation formula is as follows:

$$\mathrm{WAR} = \frac{\sum_{i=1}^{N} n_i * TP_i}{\sum_{i=1}^{N} n_i} \tag{4}$$

- **Unweighted accuracy rate**

Unweighted Accuracy Rate (UAR) simply averages the accuracy of each class without considering the number of samples in each class. This method is more suitable for situations where the number of samples in each class is similar, as it prevents any single class from having an excessive influence. The calculation formula is as follows:

$$\mathrm{UAR} = \frac{1}{N}\sum_{i}^{N} \frac{TP_i}{n_i} \tag{5}$$

C. Implementation Details

The primary experimental environment used in this study is shown in Table 1. During the network training process, the Adam optimizer was employed to effectively ensure that the training loss reaches a global optimum. Additionally, to ensure model stability and reproducibility, the random seed was set to 3047.

Table 1 Experimental Setup

Environment	Details
Operator System	Ubuntu18.04
GPU	GeForce RTX2080Ti Video memory: 11GB
GPU Acceleration platform	CUDA11.1
Framework	PyTorch 1.8 + python3.8

In the DFEW dataset, the original data provides individual face images extracted from single-label videos, and the data is divided into five groups, with each group split into training and testing sets at a ratio of approximately 8:2. Therefore, a portion of the training set is further divided to create a validation set, resulting in a training, validation, and testing ratio of approximately 6:2:2. For the self-constructed dataset, due to the limited data volume, 20% of the self-constructed data is used to fine-tune the classification layer on the model parameters trained on the DFEW dataset. The hyperparameters used during model training are shown in the Table 2.

Table 2 Training Hyperparameter Configuration

Hyperparameter	Parmeter (DFEW)	Parmeter (our dataset)
Optimizer	ADAM	ADAM
Learning Rate	1e-3	1e-5
Weight decay	1e-3	1e-5
Random seed	3047	3047
epoch	50	10
batch_size	32	32
device	Cuda 0	Cuda 0

D. Comparison With State-of-The-Art Methods

The table below shows that our model structure has significant improvements in recognizing multiple individual emotion categories as well as overall performance compared to other models. It can also be seen that the use of class-weighted optimization strategies effectively alleviates the long-tail problem caused by the uneven distribution of data samples, leading to a substantial improvement in the recognition of the Fear category. However, due to the limited amount of data in the Disgust category, the recognition performance for this category is very poor.

Table 3 Experimental Results on DFEW Dataset

Methods	Accuracy of Each Emotion(%)							Metrics(%)	
	Happy	Sadness	Neutral	Anger	Surprise	Disgust	Fear	**UAR**	**WAR**
VGG11+LSTM[24]	76.89	37.65	58.04	60.70	43.70	0.00	19.73	42.39	53.70
EC-STFL[25]	79.18	49.05	57.85	60.98	46.15	2.76	21.51	45.35	56.51
Resnet18+LSTM[29]	83.56	61.56	68.27	65.29	51.26	0.00	29.34	51.32	63.85
Resnet18+ GRU[23]	82.87	63.83	65.06	68.51	52.00	0.86	30.14	51.68	64.02
Former-DFER[23]	84.05	62.57	67.52	70.03	56.43	3.45	31.78	53.69	65.70
EST[16]	86.87	66.58	67.18	71.84	47.52	**5.52**	28.49	53.43	65.85
STT[38]	87.36	**67.90**	64.97	71.24	53.1	3.49	34.04	**54.58**	66.45
MTFNet (w/o)	**91.12**	63.23	68.33	**72.13**	**68.10**	0.00	0.00	51.83	66.80
MTFNet(Ours)	89.78	58.73	**69.79**	65.90	52.90	0.00	**37.78**	53.55	**67.59**

Due to the limited amount of data that we have collected, it is not feasible to retrain the model from scratch. Therefore, we fine-tuned the classification head of the model trained on the DFEW dataset. Table 4 below compares the recognition performance of our model with that of some other models on the self-constructed dataset.

Table 4 Comparison of Custom Group Learning Datasets

Methods	Accuracy of Each Emotion(%)				Metrics(%)	
	Confuse	Bored	Interest	Happy	**UAR**	**WAR**
EST[36]	40.00	63.63	37.78	86.67	57.00	55.56
STT[37]	40.00	72.72	**49.06**	86.67	62.13	62.62
MTFNet(Ours)	40.00	**90.91**	47.72	**93.33**	**67.85**	**64.65**

It can be seen that on the self-constructed group learning dataset, our model outperforms the EST and STT models in terms of accuracy for multiple individual categories, as well as weighted and unweighted accuracy rates. The accuracy of the model for the interest category is relatively low, mainly because some data in the interest category are similar to the confuse category, making them more challenging to recognize.

E. Ablation study

The results of the module ablation experiments on the DFEW dataset are shown in the Table 5 below:

Table 5 Results of Ablation Experiments on Modules

MFFF	FSOP	DFEW	
		WAR	UAR
		65.02	51.00
	√	65.31	52.73
√		66.35	51.22
√	√	**67.60**	**53.55**

The experimental results indicate that both the MFFF module and the FSOP module significantly contribute to this task, and their combination can further improve the model's performance.

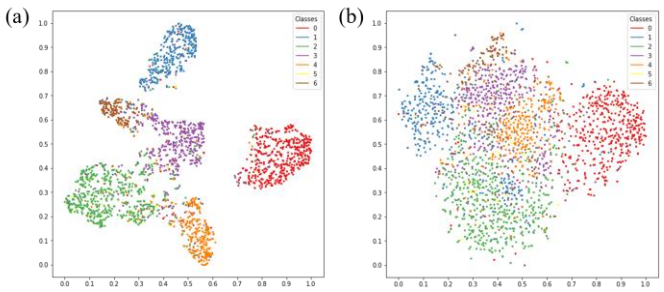

Figure 8 Visualisation of feature distribution (t-SNE)

We used t-SNE to visualize the learned feature distributions with and without the MFFF module, as shown in Figure 8. In the right figure, distinct clustering and clearer inter-class boundaries can be observed. The feature distribution obtained by our model shows greater inter-class differences and tighter intra-class cohesion, with clear boundary lines. This also verifies that the introduction of the MFFF module enriches the spatial semantic features of the images, thereby enhancing the model's accuracy.

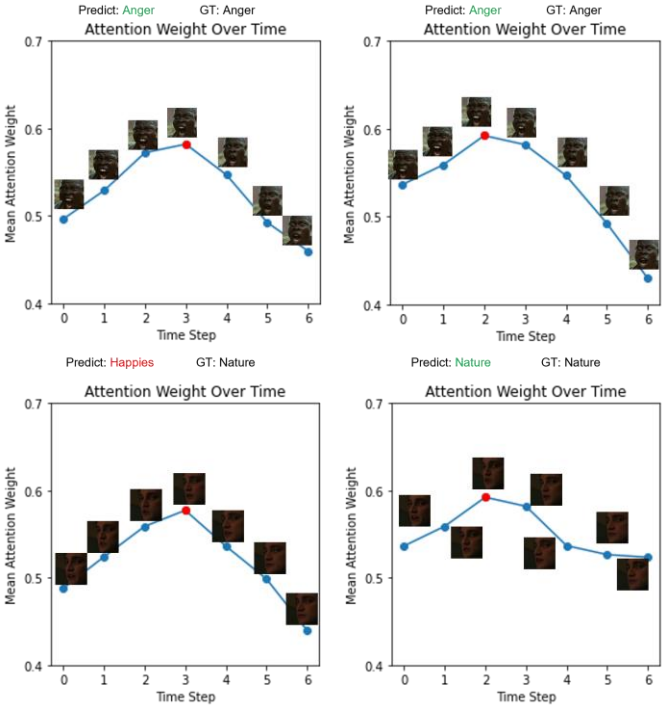

Figure 9 Temporal Attention Visualisation

We visualized the attention weights of the last layer features in the Temporal Transformer in Figure 9. The comparison shows that in the Anger sample, our model can effectively capture the differences in facial emotions at different time points, amplifying the information variance. In the Neutral sample, our model can focus more on the overall facial changes, thus not being affected by the slight movement of the mouth corners in the image. Overall, the introduction of the FSOP module reduces the impact of the multi-head attention mechanism's excessive focus on peak variation frames, leading to better temporal modeling of emotional information.

F. Limitations and Future Works

Currently, the challenges in facial expression recognition within group learning scenarios mainly stem from data scarcity and the loss of facial features in complex environments. To address this, we introduced the MFFF module in the spatial feature extraction part to capture multi-scale features and collected a small evaluation dataset. In the future, the performance of recognition in this scenario can be improved through the following approaches:

- Dataset Expansion: High-quality data will help the model better capture high-quality features.
- Semi-Supervised Learning: By combining general emotion datasets for learning, this approach can alleviate the issue of data scarcity in specific scenarios and expand the existing small dataset.

V. CONCLUSION

Given the complexity of group learning scenarios and the lack of relevant datasets, this paper proposes the MTFNet model, which is based on the efficiency of the spatial-temporal framework for recognition. Our model can be deployed on terminals and achieve real-time monitoring. In the spatial

feature extraction part, we introduced a Multi-scale Facial Feature Fusion Module (MFFF) based on GS convolution, which integrates features from different semantic spaces while controlling the number of parameters, mitigating the decline in performance caused by occlusion and uneven lighting in complex group learning scenarios. In the temporal feature extraction phase, we improved the Frame-wise Shuffle Order Prediction Module (FSOP) to guide the model's attention to changes between frames, enhancing the model's ability to capture dynamic emotional changes. Additionally, we collaborated with educational experts to collect data and construct a small evaluation dataset suitable for group learning scenarios in real settings. As a result, our model achieved excellent performance on both the DFEW dataset and the self-made group learning scenario dataset, with WAR values of 67.59% and 64.65%, respectively.

VI. ACKNOWLEDGMENT

We sincerely appreciate the generous support from several funding agencies for this research. First and foremost, this study was funded by the National Natural Science Foundation of China under the project titled "Multi-modal Assessment Models and Interventions for Student Engagement in Cooperative Learning" (No. 62277007). Additionally, we are grateful for the support from the Guangdong Province Undergraduate Course Teaching and Research Office Construction Project (No. jx2022303), as well as the aid provided by the Guangdong Province University Characteristic Innovation Project (Natural Science) (No. 2022KTSCX205). The funding from these institutions was crucial in facilitating the successful execution of our research plan and achieving the anticipated outcomes.

REFERENCES

[1] A. Mehrabian, ‘Communication without words’, in Communication theory, Routledge, 2017, pp. 193 – 200. Accessed: Jul. 05, 2024.

[2] R. Adyapady R and A. Basava, “A comprehensive review of facial expression recognition techniques,” Multimedia Systems, vol. 29, Jul. 2022, doi: 10.1007/s00530-022-00984-w.

[3] Ö. Sümer, P. Goldberg, S. D’Mello, P. Gerjets, U. Trautwein, and E. Kasneci, ‘Multimodal engagement analysis from facial videos in the classroom’, IEEE Transactions on Affective Computing, vol. 14, no. 2, pp. 1012 – 1027, 2021.

[4] Y. Yi, H. Zhang, W. Zhang, Y. Yuan, and C. Li, “Fatigue Working Detection Based on Facial Multi-Feature Fusion,” IEEE Sensors Journal, vol. PP, pp. 1 – 1, Mar. 2023, doi: 10.1109/JSEN.2023.3239029.

[5] S. Gupta, P. Kumar, and R. K. Tekchandani, “Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models,” Multimed Tools Appl, vol. 82, no. 8, pp. 11365 – 11394, Mar. 2023, doi: 10.1007/s11042-022-13558-9.

[6] R. Pekrun and L. Linnenbrink-Garcia, “Academic Emotions and Student Engagement,” in Handbook of Research on Student Engagement, S. L. Christenson, A. L. Reschly, and C. Wylie, Eds., Boston, MA: Springer US, 2012, pp. 259 – 282. doi: 10.1007/978-1-4614-2018-7_12.

[7] H. Zeng et al., “EmotionCues: Emotion-Oriented Visual Summarization of Classroom Videos,” IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 7, pp. 3168 – 3181, Jul. 2021, doi: 10.1109/TVCG.2019.2963659.

[8] H. Li, H. Niu, Z. Zhu, and F. Zhao, “Intensity-Aware Loss for Dynamic Facial Expression Recognition in the Wild,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 1, Art. no. 1, Jun. 2023, doi: 10.1609/aaai.v37i1.25077.

[9] H. Kim, J.-H. Lee, and B. C. Ko, “Facial Expression Recognition in the Wild Using Face Graph and Attention,” IEEE Access, vol. 11, pp. 59774 – 59787, 2023, doi: 10.1109/ACCESS.2023.3286547.

[10] J. Liao, Y. Lin, T. Ma, S. He, X. Liu, and G. He, “Facial Expression Recognition Methods in the Wild Based on Fusion Feature of Attention Mechanism and LBP,” Sensors, vol. 23, no. 9, Art. no. 9, Jan. 2023, doi: 10.3390/s23094204.

[11] R. Zhi, M. Flierl, Q. Ruan, and W. Kleijn, “Graph-Preserving Sparse Nonnegative Matrix Factorization With Application to Facial Expression Recognition,” IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society, vol. 41, pp. 38 – 52, Feb. 2011, doi: 10.1109/TSMCB.2010.2044788.

[12] T. Ojala, M. Pietikainen, and T. Maenpaa, ‘Multiresolution gray-scale and rotation invariant texture classification with local binary patterns’, IEEE Transactions on pattern analysis and machine intelligence, vol. 24, no. 7, pp. 971 – 987, 2002.

[13] M. Shafiq and Z. Gu, ‘Deep residual learning for image recognition: A survey’, Applied Sciences, vol. 12, no. 18, p. 8972, 2022.

[14] S. Hochreiter and J. Schmidhuber, ‘Long Short-Term Memory’, Neural Computation, vol. 9, no. 8, pp. 1735 – 1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. of Advances in Neural Information Processing Systems (NIPS), pp. 5998-6008, Dec. 2017.

[16] Y. Liu, W. Wang, C. Feng, H. Zhang, Z. Chen, and Y. Zhan, ‘Expression snippet transformer for robust video-based facial expression recognition’, Pattern Recognition, vol. 138, p. 109368, Jun. 2023, doi: 10.1016/j.patcog.2023.109368.

[17] G. Zhao, Y. Zhang, and J. Chu, ‘A multimodal teacher speech emotion recognition method in the smart classroom’, Internet of Things, vol. 25, p. 101069, 2024.

[18] M. Li et al., ‘Multimodal Emotion Recognition and State Analysis of Classroom Video and Audio Based on Deep Neural Network’, Journal of Interconnection Networks, Feb. 2022, doi: 10.1142/S0219265921460117

[19] Z. Zhu, X. Zheng, T. Ke, and G. Chai, ‘Emotion Recognition in Learning Scenes Supported by Smart Classroom and Its Application.’, Traitement du Signal, vol. 40, no. 2, 2023, Accessed: Jul. 05, 2024.

[20] E. Yadegaridehkordi, N. F. B. M. Noor, M. N. B. Ayub, H. B. Affal, and N. B. Hussin, ‘Affective computing in education: A systematic review and future research’, Computers &

Education, vol. 142, p. 103649, Dec. 2019, doi: 10.1016/j.compedu.2019.103649.

[21] P. Ekman and W. V. Friesen, ‘Constants across cultures in the face and emotion’, Journal of Personality and Social Psychology, vol. 17, no. 2, pp. 124 – 129, 1971, doi: 10.1037/h0030377.

[22] P. Ekman and W. V. Friesen, Facial action coding system: investigator’s guide. Palo Alto, Calif.: Consulting Psychologists Press, 1978.

[23] Z. Zhao and Q. Liu, ‘Former-DFER: Dynamic Facial Expression Recognition Transformer’, in Proceedings of the 29th ACM International Conference on Multimedia, New York, NY, USA, Oct. 2021, pp. 1553 – 1561. doi: 10.1145/3474085.3475292.

[24] Y. Wang et al., ‘FERV39k: A Large-Scale Multi-Scene Dataset for Facial Expression Recognition in Videos’, 2022, pp. 20922 – 20931. Accessed: Apr. 10, 2024.

[25] X. Jiang et al., ‘DFEW: A Large-Scale Database for Recognizing Dynamic Facial Expressions in the Wild’, in Proceedings of the 28th ACM International Conference on Multimedia, New York, NY, USA, Oct. 2020, pp. 2881 – 2889. doi: 10.1145/3394171.3413620.

[26] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, ‘Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron’, in Proceedings Third IEEE International Conference on Automatic face and gesture recognition, 1998, pp. 454 – 459. Accessed: Jul. 07, 2024.

[27] C. Shan, S. Gong, and P. W. McOwan, ‘Robust facial expression recognition using local binary patterns’, in IEEE International Conference on Image Processing 2005, 2005, vol. 2, p. II – 370. Accessed: Jul. 07, 2024.

[28] J. Whitehill and C. W. Omlin, ‘Haar features for FACS AU recognition’, in 7th international conference on automatic face and gesture recognition (FGR06), 2006, pp. 5-pp. Accessed: Jul. 07, 2024.

[29] M. F. Valstar et al., ‘Fera 2015-second facial expression recognition and analysis challenge’, in 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015, vol. 6, pp. 1 – 8. Accessed: Jul. 07, 2024.

[30] V. Vielzeuf, S. Pateux, and F. Jurie, ‘Temporal multimodal fusion for video emotion classification in the wild’, in Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow UK, Nov. 2017, pp. 569 – 576. doi: 10.1145/3136755.3143011.

[31] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, ‘Empirical evaluation of gated recurrent neural networks on sequence modeling’, CoRR, vol. abs/1412.3555, 2014

[32] Y. Fan, X. Lu, D. Li, and Y. Liu, ‘Video-based emotion recognition using CNN-RNN and C3D hybrid networks’, in Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo Japan, Oct. 2016, pp. 445 – 450. doi: 10.1145/2993148.2997632.

[33] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, ‘Learning spatiotemporal features with 3d convolutional networks’, in Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489 – 4497. Accessed: Jul. 07, 2024.

[34] L. Sun, Z. Lian, B. Liu, and J. Tao, ‘MAE-DFER: Efficient Masked Autoencoder for Self-supervised Dynamic Facial Expression Recognition’, in Proceedings of the 31st ACM International Conference on Multimedia, Ottawa ON Canada, Oct. 2023, pp. 6110 – 6121. doi: 10.1145/3581783.3612365.

[35] P. L. Wang, Q. C. Ke, and J. Q. Zhang, ‘Research on application of brain-computer interface in smart classroom’, Open Education Research, vol. 26, pp. 72 – 81, 2020.

[36] S. Bo, L. Yongna, C. Jiubing, L. Jihong, and Z. Di, ‘Emotion analysis based on facial expression recognition in smart learning environment’, Modern Distance Education Research, vol. 2, pp. 96 – 103, 2015.

[37] Z. H. Zhan, ‘An emotional and cognitive recognition model for distance learners based on intelligent agent-the coupling of eye tracking and expression recognition techniques’, Mod. Dist. Educ. Res, vol. 5, pp. 100 – 105, 2013.

[38] F. Ma, B. Sun, and S. Li, ‘Spatio-Temporal Transformer for Dynamic Facial Expression Recognition in the Wild’. arXiv, May 10, 2022. doi: 10.48550/arXiv.2205.04749.

[39] H. Li, J. Li, H. Wei, Z. Liu, Z. Zhan, and Q. Ren, “Slim-neck by GSConv: A lightweight-design for real-time detector architectures,” J Real-Time Image Proc, vol. 21, no. 3, p. 62, Jun. 2024, doi: 10.1007/s11554-024-01436-6.