xLSTM-FER: Enhancing Student Expression Recognition with Extended Vision Long Short-Term Memory Network

Qionghao Huang(✉) and Jili Chen

Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, Jinhua, Zhejiang, China
{qhhuang,irelia}@zjnu.edu.cn

Abstract. Student expression recognition has become an essential tool for assessing learning experiences and emotional states. This paper introduces xLSTM-FER, a novel architecture derived from the Extended Long Short-Term Memory (xLSTM), designed to enhance the accuracy and efficiency of expression recognition through advanced sequence processing capabilities for student facial expression recognition. xLSTM-FER processes input images by segmenting them into a series of patches and leveraging a stack of xLSTM blocks to handle these patches. xLSTM-FER can capture subtle changes in real-world students' facial expressions and improve recognition accuracy by learning spatial-temporal relationships within the sequence. Experiments on CK+, RAF-DF, and FERplus demonstrate the potential of xLSTM-FER in expression recognition tasks, showing better performance compared to state-of-the-art methods on standard datasets. The linear computational and memory complexity of xLSTM-FER make it particularly suitable for handling high-resolution images. Moreover, the design of xLSTM-FER allows for efficient processing of non-sequential inputs such as images without additional computation.

Keywords: Facial Expression Recognition · Student Academic Performance · Memory Network · Vision xLSTM

1 Introduction

Student facial expression recognition is a burgeoning field with significant implications for educational technology. By analyzing students' facial cues, educators can gain insights into their emotional states, engagement levels [25], cognitive load [12], and academic performance [8,11] during learning activities [9]. The current student face recognition systems primarily include those based on traditional CNN-based and Vision Transformer [5] (ViT)-based approaches. The lightweight and efficient characteristics of CNNs have attracted the attention of early education researchers, leading to the development of a series of face recognition systems and teaching environments based on CNNs [22]. The ViT has

W. Zhang et al. (Eds.): APWeb-WAIM 2024, CCIS 2246, pp. 249–259, 2025.
https://doi.org/10.1007/978-981-96-0055-7_21

replaced CNN as a more robust backbone network for student facial expression recognition. The Vision Transformer, leveraging self-attention for global image modeling, has surpassed the performance of CNNs in both teaching feedback systems [28] and the assessment of learning outcomes [10,29].

However, for CNNs, the main limitation is their lack of global receptive fields and dynamic weighting capabilities, which can restrict their ability to capture long-range dependencies and integrate information from the entire image [13]. Besides, this advantage of ViTs comes at the cost of quadratic complexity in terms of image sizes, which leads to a significant computational overhead when dealing with dense prediction tasks such as object detection and semantic segmentation [33].

To address the aforementioned issues, we propose the xLSTM-FER. xLSTM-FER begins by segmenting the input image into a series of non-overlapping patches, converting the 2D image into a 1D token sequence with added learnable 2D positional encodings to retain spatial information. These sequences are then fed into an xLSTM encoder composed of stacked xLSTM blocks. The xLSTM blocks maintain a linear complexity while capturing long-range dependencies and spatial-temporal dynamics within the image sequence. Each xLSTM block employs a modified LSTM layer (mLSTM) that uses matrix values for memory retrieval, enhancing the model's capacity to discern subtle facial movements. To overcome the inherent difficulty of parallel processing in LSTM, the mLSTM utilizes a memory matrix to enhance parallel capabilities. By integrating different path traversals, the model achieves a comprehensive image representation. The summary of our contributions is as follows:

- We propose xLSTM-FER, which segments input images into a series of patches and processes them through a stack of xLSTM blocks, allowing the model to capture subtle facial expression changes and improve recognition accuracy by learning the spatial-temporal dynamics within the sequence.
- The xLSTM-FER has the capabilities of parallelization and scalability through the memory matrix calculation. With its linear computational and memory complexity, which is essential for capturing clear and detailed student expressions and making xLSTM-FER a more practical solution for real-world applications.
- The extensive empirical evaluations of the xLSTM-FER model on multiple standard datasets demonstrate its superior performance in facial expression recognition tasks, including a perfect score on the CK+ [18] dataset, and shows substantial improvements over previous state-of-the-art methods on both RAF-DB [16] and FERplus [2] datasets.

2 Related Work

2.1 Student Facial Expression Recognition in Learning Environment

Early work utilizes Convolutional Neural Networks (CNNs) as the backbone for facial expression recognition tasks. Mohamad *et al.* [21] use a VGG-B network to calculate the level of student engagement in MOOCs based on their

facial expressions. Lasri *et al.* [15] demonstrate a CNN-based automatic facial recognition system in educational settings can assist teachers in adjusting their teaching strategies and materials according to the emotional responses of students. Wang *et al.* [26] introduce a framework integrating an enhanced MobileViT [19] model with an online platform for real-time student emotion analysis. To analyze student expressions and inform teaching strategies, Ling *et al.* [17] present a classroom-based FER system using YOLO and ViT. The computational demands of ViTs grow quadratically with the self-attention mechanism, which can be prohibitive for applications requiring high-resolution processing. To make facial recognition more efficient in educational scenarios, xLSTM-FER demonstrates linear computational and memory complexity, making it more suitable for training and practical deployment.

2.2 Long Short-Term Memory Network

LSTM [7] is a type of recurrent neural network (RNN) architecture that is particularly good at learning order dependence in sequence prediction problems. Recently, Beck [3] propose improvements to LSTM, including exponential gating and novel memory structures, to address the limitations of LSTM and enable it to scale to larger model sizes. Alkin *et al.* [1] verify that xLSTM is also applicable as a visual backbone network. Compared to CNNs, xLSTM has the characteristic of being scalable, and compared to Vision Transformers, it has a more linear complexity which makes it easier to deploy in practice. However, its application in student facial expression recognition remains unexplored. Therefore, we propose xLSTM-FER to explore the potential application of LSTM-based models and overcome the challenge in student expression recognition.

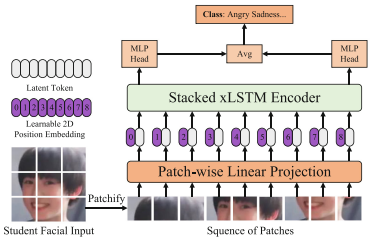
3 Methodology

3.1 Patch Embedding

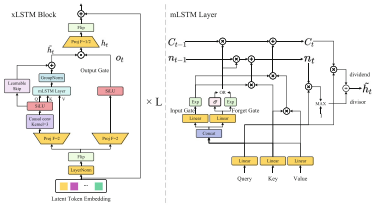
The overall architecture of our network is shown in Fig. 1a. We first perform patchification on the image. The input image $x \in \mathrm{R}^{H \times W \times C}$ is divided into a grid of non-overlapping patches. Each patch is a small square or rectangle of pixels with a width of P. Then, each patch is flattened into a sequence of pixel values $X_p \in \mathrm{R}^{N \times (P^2 \times C)}$, where $N = HW/(P^2)$. The flattened patch sequences are then linearly projected to a higher-dimensional space. To provide the model with information about the relative positions of the patches, we add learnable 2D positional embeddings to the patch sequences.

3.2 xLSTM Encoder

xLSTM Block. The xLSTM encoder is a structure composed of L-layer stacked xLSTM Blocks, as shown in Fig. 1b. The xLSTM Block begins by layer-normalizing and then inverting the input. One branch doubles the channels



(a) The pipeline of our xLSTM-FER.



(b) Left: The model diagram of the xLSTM Block. Right: The model diagram of the xLSTM Layer.

Fig. 1. Framework of our xLSTM-FER.

(F = 2) to construct the output gating, while the other branch uses a causal convolution layer (Kernel = 3) to build the input for the mLSTM layer, which includes the query and key branch vectors for the linear attention mechanism [14], with the value vector bypassing the causal convolution. The output of the mLSTM layer is passed through a group normalization layer [30] and is then summed with the output of the causal convolution via a weighted residual connection to obtain $\tilde{h}_t$, and $\tilde{h}_t$ is gated with the result of the output gate o_t to obtain the output of the hidden state h_t. Finally, the channels are halved (F = 1/2) and sum with the input embeddings of the block through a residual connection to obtain the entire xLSTM block's output. This high-capacity storage capability enables the model to distinguish between subtle differences in facial expressions, crucial for identifying even the most nuanced emotions. This scalability is essential for creating robust systems capable of operating in diverse real-world environments.

mLSTM Layer. The mLSTM employs a FlashAttention mechanism, which is simulated using query, key, and value to guide the updates of both the cell state and the normalizer state, and subsequently outputs the results of the hidden layer as illustrated in Fig. 1b. Specifically, the mLSTM layer first performs linear projections on the query, key, and value vectors:

$$\begin{aligned} \text{QueryMapping} \quad q_t &= W_q\, x_q + b_q, \\ \text{ScaledKeyMapping} \quad k_t &= \frac{1}{\sqrt{d}} W_k\, x_k + b_k, \\ \text{ValueMapping} \quad v_t &= W_v\, x_v + b_v, \end{aligned} \tag{1}$$

where x_q, x_k, and x_v, represent the input query, key, and value vectors respectively, while W_q, W_k, and W_v are the corresponding mapping matrices. b_q, b_k, and b_v are the corresponding bias terms. By concatenating the mapped query, key, and value vectors, the input x_t is obtained for the memory network to perform memory updates. The xLSTM uses an input gate and a forget gate to control the situation of memory updates and employs exponential gating and OR gating to facilitate the matrix memory calculation:

$$\begin{aligned} \text{Input Gate:} \quad & i_t = \exp(\tilde{i}_t), \quad \tilde{i}_t = w_i^\top x_t + b_i, \\ \text{Forget Gate:} \quad & f_t = \exp(\tilde{f}_t) \,\text{OR}\, \sigma(f_t), \quad \tilde{f}_t = w_f^\top x_t + b_f, \end{aligned} \tag{2}$$

where $w_i^\top$, $w_f^\top$, b_i, b_f denote the weight vectors and bias terms corresponding to the input gate and forget gate, respectively. The σ denotes the activation function, and $\exp(\cdot)$ signifies the exponential operation. The mLSTM expands the memory cell into a matrix. By integrating the update mechanism of LSTM with the information retrieval scheme from Transformers, mLSTM introduces an attention-integrated cell state and hidden state update scheme, enabling the extraction of memories from different time steps:

$$\begin{aligned} \text{Cell State:} \quad & C_t = f_t C_{t-1} + i_t v_t k_t^\top, \\ \text{Normalizer State:} \quad & n_t = f_t n_{t-1} + i_t k_t, \\ \text{Output Gate:} \quad & o_t = \sigma(\tilde{o}_t), \quad \tilde{o}_t = W_o x_t + b_o, \\ \text{Hidden State:} \quad & h_t = o_t \odot \tilde{h}_t, \quad \tilde{h}_t = C_t q_t / \max\{|n_t^\top q_t|, 1\}, \end{aligned} \tag{3}$$

inspired by [23], the cell state uses a weighted sum according to proportions, where the forget gate corresponds to the weighted proportion of memory, and the input gate corresponds to the weighted proportion of the key-value pair to satisfy the covariance-based update rule. The mLSTM employs a normalizer that weights key vectors. Ultimately, through normalization and weighted control by the output gate, the hidden state h_t of the network is obtained. The mLSTMs employ matrix values to process memory retrieval, which allows the retrieval process in mLSTMs to be conducted directly through matrix multiplication. The hidden state from timestep $t-1$ is not included in the processing flow, which greatly enhances the parallelization capability of the mLSTM. The mLSTM introduction of matrix memory and parallelization brings a new level of sophistication to facial expression recognition systems. By employing a matrix memory cell, the mLSTM can store a richer feature representation, capturing the intricate details and variations that define different emotional expressions. Moreover, the parallelization feature of the mLSTM block enables the model to process this complex facial data more efficiently, significantly reduce the computational load.

3.3 Path Transfer

By integrating the outcomes from these various views [24], a more accurate modeling of the sequence can be achieved. Traditional sequence modeling typically has two path traversal schemes: forward traversal and backward traversal. We have integrated four path scanning schemes: forward and backward bidirectional in the column direction and forward and backward bidirectional in the row direction. The xLSTM incorporates a flip module to achieve a more comprehensive image representation by weighting four paths of the image data.

3.4 Classification Head

The output of the xLSTM module will be mapped to the classification dimensions. The current main methods of token aggregation are as follows: 1. Using a learnable [CLS] token placed at the beginning [5] or middle [34] of the sequence. 2. Applying average pooling to the entire sequence. 3. Using the average of the first token and the last token as the input for the classification head. In the vast majority of datasets, objects are typically centered around the middle token by default. To avoid this bias and enhance the generality of our model, our experiments adopt the last scheme mentioned. Our loss function is the cross-entropy loss function:

$$\mathcal{L} = -\sum_{n=1}^{N} y \log(\hat{y}) \tag{4}$$

4 Experiments

4.1 Datasets and Metrics

We conduct experiments on three datasets in FER research: CK+ [18], RAF-DB [16], and FERplus [2]. We report the Top-1 accuracy on the seven-category task as the evaluation metric. Here is a brief introduction to the datasets. **CK+.** The CK+ dataset includes annotations for the following emotions: Anger, Contempt, Disgust, Fear, Happy, Sadness, and Surprise. The CK+ dataset comprises 784 training samples and 197 test samples. **RAF-DB.** The RAF-DB encompasses seven basic emotional categories: surprise, fear, disgust, happiness, sadness, anger, and neutrality. The training subset encompasses 12,271 images, while the test subset consists of 3,068 images. **FERplus.** The FERplus dataset is an enhanced version of the original FER dataset. The FERplus dataset categorizes expressions into eight distinct emotions: anger, disgust, fear, happy, sad, surprise, neutral, and contempt. The dataset comprises a total of 28,709 images for training, along with 3,589 images allocated for validation and 3,589 designated for testing purposes.

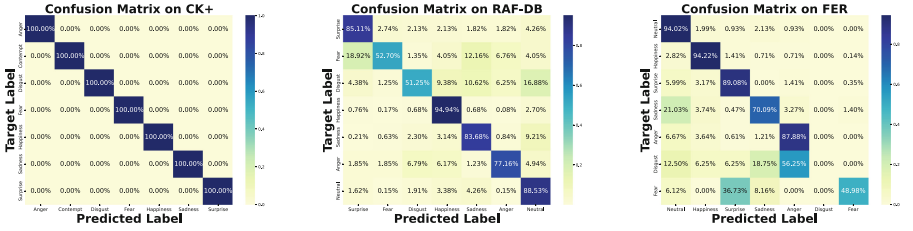
4.2 Experiment Settings

We conduct experiments with a patch size set to 16×16, the number of stacked xLSTM layers being 26, and the base dimension of the model being 384, which means the dimensions of the query, key, and value vectors are 768. The number of our attention heads is 192.

4.3 Results

Table 1. Results on CK+, RAF-DB, and FERplus. The previous state-of-the-art (SOTA) values are marked with underlines, while the current SOTA values are marked in bold. All reported values are based on the **"from scratch"** setting.

Method	CK+	RAF-DB	FERplus
FER-GCN [6]	99.54%	-	-
FMPN [4]	98.06%	-	-
FAN [20]	99.70%	-	-
SelfCureNet [27]	-	78.31%	83.42%
ViT [5]	96.88%	63.75%	73.36%
MA-Net [32]	-	67.48%	-
EAC [31]	-	73.73%	75.77%
xLSTM-FER(ours)	**100%**	**87.06%**	**88.94%**
Rank	**1**	**1**	**1**



(a) Confusion Matrix on CK+.

(b) Confusion Matrix on RAF-DB

(c) Confusion Matrix on FERplus.

Fig. 2. Confusion Metrics on Datasets.

We compare xLSTM-FER with the recent CNN-based models such as FER-GCN [6], EAC [31] and ViT-based face recognition models including ViT [5], MA-Net [32], and others to verify the effectiveness of xLSTM-FER. All experiments are conducted from scratch. The experimental results are shown in Table 1.

Results on CK+. In the CK+ dataset, the outcomes presented in Table 1 reveal that our technique, xLSTM-FER, has pioneered a perfect accuracy rate of 100% for classifying facial expressions. The confusion matrix in Fig. 2a indicates that xLSTM-FER has achieved 100% accuracy across all categories. These outcomes excel over the sota methods in the realms of both video and image-based facial expression analysis. Because xLSTM-FER successfully captures the interdependencies among different patch blocks.

Results on RAF-DB. xLSTM-FER achieves an impressive overall accuracy of 87.06% and shows a 14% improvement over the previous sota values, demonstrating superior performance compared to other models. In contrast, ViT only achieves a lower accuracy of 63.75%, and another visual transformer-based FER model, MA-Net, does not perform well on this in-the-wild dataset. xLSTM-FER achieves a competitive accuracy of 87.06%, indicating its robust performance compared to current state-of-the-art methods.

Results on FERplus. On the FERplus dataset, our model has outperformed all contemporary methods, attaining an accuracy rate of 88.94%. xLSTM-FER shows a 4.5% improvement compared with previous sota values. This confirms that the synergistic effect of the memory gating and attention mechanisms within xLSTM-FER can achieve an accurate representation of facial images.

4.4 Case Analysis

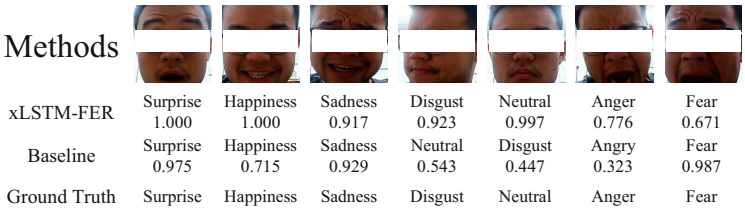


Fig. 3. A case study on the accuracy of xLSTM-FER compared to the baseline model (EAC [31]) in real-world examples.

To further verify the advantages of xLSTM-FER over the baseline, we test several photos in the learning environment. The test results are shown in Fig. 3. We find that, except for the "Fear" and "Sadness" categories, xLSTM-FER can provide more accurate predictions with higher confidence compared to EAC in other categories. This indicates that the memory network of xLSTM in its image extraction approach can adapt to the real-world needs of students' FER tasks even with the linear complexity.

5 Conclusion

To overcome the quadratic complexity in traditional student facial expression recognition, this paper presents xLSTM-FER, which has profound implications for the assessment of learning experiences and emotional states. The innovative approach of xLSTM-FER in segmenting input images into patches and processing them through a stack of xLSTM blocks. Our experimental results

on CK+, RAF-DB, and FERplus not only validate the potential of xLSTM-FER in student facial expression recognition tasks but also highlight its competitive performance when compared to state-of-the-art methods on standard datasets. The linear computational and memory complexity of xLSTM-FER is a significant advantage, making it exceptionally well-suited for processing high-resolution images, which is essential for the clear and detailed capture of student expressions. We are confident that with further optimization and fine-tuning, xLSTM-FER will evolve as a significant tool in student expression recognition.

Acknowledgement. The research project is supported by the National Natural Science Foundation of China (No. 62207028), partially by Zhejiang Provincial Natural Science Foundation (No. LY23F020009), and the Key R&D Program of Zhejiang Province (No. 2022C03106), and Scientific Research Fund of Zhejiang Provincial Education Department (No. 2023SCG367).

References

1. Alkin, B., Beck, M., Pöppel, K., Hochreiter, S., Brandstetter, J.: Vision-LSTM: xLSTM as generic vision backbone. arXiv preprint arXiv:2406.04303 (2024)
2. Barsoum, E., Zhang, C., Ferrer, C.C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 279–283 (2016)
3. Beck, M., et al.: xLSTM: extended long short-term memory. arXiv preprint arXiv:2405.04517 (2024)
4. Chen, Y., Wang, J., Chen, S., Shi, Z., Cai, J.: Facial motion prior networks for facial expression recognition. In: 2019 IEEE Visual Communications and Image Processing (VCIP), pp. 1–4. IEEE (2019)
5. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Fan, Y., Lam, J.C., Li, V.O.: Video-based emotion recognition using deeply-supervised neural networks. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp. 584–588 (2018)
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
8. Huang, Q., Chen, J.: Enhancing academic performance prediction with temporal graph networks for massive open online courses. J. Big Data **11**(1), 52 (2024)
9. Huang, Q., Huang, C., Huang, J., Fujita, H.: Adaptive resource prefetching with spatial-temporal and topic information for educational cloud storage systems. Knowl.-Based Syst. **181**, 104791 (2019)
10. Huang, Q., Huang, C., Wang, X., Jiang, F.: Facial expression recognition with grid-wise attention and visual transformer. Inf. Sci. **580**, 35–54 (2021)
11. Huang, Q., Zeng, Y.: Improving academic performance predictions with dual graph neural networks. Complex Intell. Syst. 1–19 (2024)
12. Jagadeesh, M., Baranidharan, B.: Facial expression recognition of online learners from real-time videos using a novel deep learning model. Multimedia Syst. **28**(6), 2285–2305 (2022)
13. Jiang, F., et al.: Face2nodes: learning facial expression representations with relation-aware dynamic graph convolution networks. Inf. Sci. **649**, 119640 (2023)

14. Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are RNNs: fast autoregressive transformers with linear attention. In: International Conference on Machine Learning, pp. 5156–5165. PMLR (2020)
15. Lasri, I., Solh, A.R., El Belkacemi, M.: Facial emotion recognition of students using convolutional neural network. In: 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS), pp. 1–6. IEEE (2019)
16. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2852–2861 (2017)
17. Ling, X., Liang, J., Wang, D., Yang, J.: A facial expression recognition system for smart learning based on yolo and vision transformer. In: Proceedings of the 2021 7th International Conference on Computing and Artificial Intelligence, pp. 178–182 (2021)
18. 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)
19. Mehta, S., Rastegari, M.: Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178 (2021)
20. Meng, D., Peng, X., Wang, K., Qiao, Y.: Frame attention networks for facial expression recognition in videos. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 3866–3870. IEEE (2019)
21. Mohamad Nezami, O., Dras, M., Hamey, L., Richards, D., Wan, S., Paris, C.: Automatic recognition of student engagement using deep learning and facial expression. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 273–289. Springer (2020)
22. Ozdemir, M.A., Elagoz, B., Alaybeyoglu, A., Sadighzadeh, R., Akan, A.: Real time emotion recognition from facial expressions using CNN architecture. In: 2019 Medical Technologies Congress (TIPTEKNO), pp. 1–4. IEEE (2019)
23. Schlag, I., Irie, K., Schmidhuber, J.: Linear transformers are secretly fast weight programmers. In: International Conference on Machine Learning, pp. 9355–9366. PMLR (2021)
24. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Signal Process. **45**(11), 2673–2681 (1997)
25. Tonguç, G., Ozkara, B.O.: Automatic recognition of student emotions from facial expressions during a lecture. Comput. Educ. **148**, 103797 (2020)
26. Wang, J., Zhang, Z.: Facial expression recognition in online course using light-weight vision transformer via knowledge distillation. In: Pacific Rim International Conference on Artificial Intelligence, pp. 247–253. Springer (2023)
27. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6897–6906 (2020)
28. Wang, K., Cheng, M.: Teaching feedback system based on VIT expression recognition in distance education. In: 2024 13th International Conference on Educational and Information Technology (ICEIT), pp. 93–97. IEEE (2024)
29. Wu, X., et al.: FER-CHC: facial expression recognition with cross-hierarchy contrast. Appl. Soft Comput. **145**, 110530 (2023)
30. Wu, Y., He, K.: Group normalization. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)

31. Zhang, Y., Wang, C., Ling, X., Deng, W.: Learn from all: erasing attention consistency for noisy label facial expression recognition. In: European Conference on Computer Vision, pp. 418–434. Springer (2022)
32. Zhao, Z., Liu, Q., Wang, S.: Learning deep global multi-scale and local attention features for facial expression recognition in the wild. IEEE Trans. Image Process. **30**, 6544–6556 (2021)
33. Zhou, S., Wu, X., Jiang, F., Huang, Q., Huang, C.: Emotion recognition from large-scale video clips with cross-attention and hybrid feature weighting neural networks. Int. J. Environ. Res. Public Health **20**(2), 1400 (2023)
34. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)