World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3060

INTELLIGENT RECOGNITION OF STUDENTS' LEARNING STATES THROUGH MICRO-EXPRESSIONS

JuanYang[1], Ling Ma[2*], XianBin Zhang[2], SiYi Tian[1]
[1]*School of Journalism and Law, Wuchang Shouyi University, Wuhan 430064, Hubei, China.*
[2]*School of Information Science and Engineering, Wuchang Shouyi University, Wuhan 430064, Hubei, China.*
[*]*Corresponding Author: Ling Ma*

Abstract: Students' facial micro-expressions can reflect their learning states, and the intelligent recognition of micro-expressions is of great significance for monitoring these states. This paper proposes a micro-expression recognition method based on a Residual Network for detecting the learning and psychological states of college students. Using ResNet18 as the backbone network, the Efficient Channel Attention (ECA) mechanism is embedded to adaptively adjust channel weights, enhancing feature representation capability. Simultaneously, the Mish activation function and Dropout layer are introduced to optimize gradient flow and reduce the risk of overfitting. During training, the Label Smoothing Cross Entropy loss function, the AdamW optimizer combined with the OneCycleLR learning rate scheduler, and an early stopping mechanism are adopted, effectively improving the model's generalization ability and training efficiency on small datasets. Experiments are based on a self-built dataset (including positive, neutral, and negative expressions, totaling 1800 grayscale images). Through data augmentation and ten-fold cross-validation, the model achieves an accuracy of 97.50%. The experimental results show that this method possesses high accuracy and robustness in micro-expression recognition tasks, providing an effective tool for monitoring the psychological states of college students and optimizing classroom teaching.
Keywords: Deep learning; Convolutional neural network; Micro-expression recognition; Learning state monitoring

1 INTRODUCTION

Currently, a significant proportion of college students face learning and psychological issues. According to the "2022 Report on the Mental Health Status of College Students" jointly released by the Institute of Psychology, Chinese Academy of Sciences and Social Sciences Academic Press, 78.52% of students showed no risk of depression, while only about half were free from anxiety risk (54.72%). Most college students reported relatively high life satisfaction, with 74.10% reaching a level of "basically satisfied" or higher. Therefore, identifying students with psychological problems is crucial for student management and psychological counseling.

Human facial expressions contain rich emotional characteristics and psychological processes, directly conveying emotional and psychological states, making them the most effective non-verbal method for emotional expression. In learning contexts, learners' facial expressions can reflect not only their emotions but also their psychological states[1]. Correctly identifying a learner's expressions allows for accurate capture of their learning emotions and states. Consequently, research and application of facial expression recognition have gradually become a hotspot in the field of intelligent education, and recognizing learners' facial expressions has become a primary method for analyzing their learning emotions and states[2].

In 1971, psychologist Ekman proposed six basic types of facial expressions: Happiness, Sadness, Fear, Anger, Surprise, and Disgust[3]. Later, Neutral was also included as a basic facial expression, forming the seven basic facial expression categories still used today.

In the early stages of expression recognition research, efforts primarily relied on handcrafted feature design or shallow learning models. However, as application scenarios shifted from controlled laboratory environments to complex real-world settings, the limitations of traditional methods in feature representation became apparent. With the rapid development of computer vision and machine learning techniques, neural networks have gradually become the main solution for expression recognition in complex scenarios due to their powerful non-linear feature extraction capabilities. Currently, deep learning technologies are increasingly applied in the field of expression recognition. Models such as Convolutional Neural Networks (CNN), Deep Belief Networks (DBN), and Recurrent Neural Networks (RNN) have become research hotspots and core technologies in this area, owing to their advantages in handling spatial image features, multi-layer semantics, and sequential dynamic information.

Facial expression recognition primarily involves three steps: preprocessing, feature extraction, and classification. Preprocessing typically involves correcting acquired images, performing face detection, followed by grayscale conversion (reducing channels, lowering complexity, and visualizing image features), cropping, scaling, translation, normalization, and other data augmentation techniques[4]. In the feature extraction and classification stage, traditional methods and deep learning-based methods are the main approaches. Among traditional methods, one proposed algorithm for facial micro-expression recognition based on hybrid features and information entropy uses 2D Gabor wavelet transform to extract micro-expression features from the eyes and eyebrows, and then uses information entropy to extract micro-expression features from the nose and mouth to improve recognition efficiency[5]. Qiao Guifang et al.

[6] proposed an expression recognition method combining Principal Component Analysis and Linear Discriminant Analysis with KPCANet-LDA. These traditional methods require specific expertise in feature design, are significantly influenced by subjective factors, and are prone to losing feature information, thereby affecting recognition efficiency. Compared to traditional recognition methods, deep learning methods hold more advantages. Chen Tuo et al. proposed a facial expression recognition network integrating spatiotemporal features, which robustly analyzes and understands spatial and temporal information of facial expressions in video sequences, effectively improving facial expression recognition performance[7]. Yang et al. proposed using a de-expression learning procedure called De-expression Residue Learning (DeRL) to extract information from expression components for facial expression recognition, thereby improving the efficiency of face recognition algorithms[8]. These deep learning methods exhibit good recognition efficiency and accuracy, but are challenging to train in practice due to the large number of parameters comprising the datasets.

Based on the above analysis, this paper proposes a method for recognizing learning states through micro-expressions based on a Residual Network. Using ResNet18 as the backbone network, the Efficient Channel Attention (ECA) mechanism is embedded to adaptively adjust channel weights and enhance feature representation capability[9-10]. Simultaneously, regularization is enhanced by introducing Dropout layers to effectively reduce overfitting risk. Furthermore, the Mish activation function replaces the traditional ReLU, further optimizing gradient flow and model performance[11]. This method significantly improves the accuracy and generalization ability of facial expression recognition while maintaining low computational complexity, making it particularly suitable for single-channel input and small classification tasks. During training, Label Smoothing Cross Entropy Loss is used to alleviate overfitting tendencies on small datasets. Additionally, a combined optimizer strategy of AdamW + OneCycleLR for super-convergence, automatic learning rate range testing and cyclical adjustment, combined with mixed-precision training and early stopping mechanisms, is employed to achieve facial expression recognition and psychological state detection on the self-built small dataset.

2 MICRO-EXPRESSION RECOGNITION AND LEARNING STATE DETECTION

2.1 Model Structure Improvements

ResNet18, as a classic deep residual network, effectively mitigates the vanishing gradient problem in deep networks through residual connections. Its 18-layer depth achieves an excellent balance between computational efficiency and feature extraction capability, making it particularly suitable for processing visual data with subtle changes like micro-expressions. However, the traditional ResNet18 has limitations in capturing low-intensity, transient micro-expressions. To address this, this study optimizes the model through feature enhancement and attention mechanisms to improve the accuracy and robustness of micro-expression recognition and the classification of students' learning-related emotions and psychological states.

Regarding the model architecture, the Efficient Channel Attention (ECA) mechanism is first embedded to enhance the response of key features by adaptively adjusting channel weights, precisely capturing local dynamic changes in micro-expressions while maintaining very low parameter count and computational overhead. Secondly, the Mish activation function replaces the traditional ReLU, alleviating gradient sparsity issues through its continuously differentiable nature and enhancing the model's non-linear expressive capability. Additionally, Dropout layers are introduced before the fully connected layer for regularization, effectively reducing overfitting risk(Figure 1).

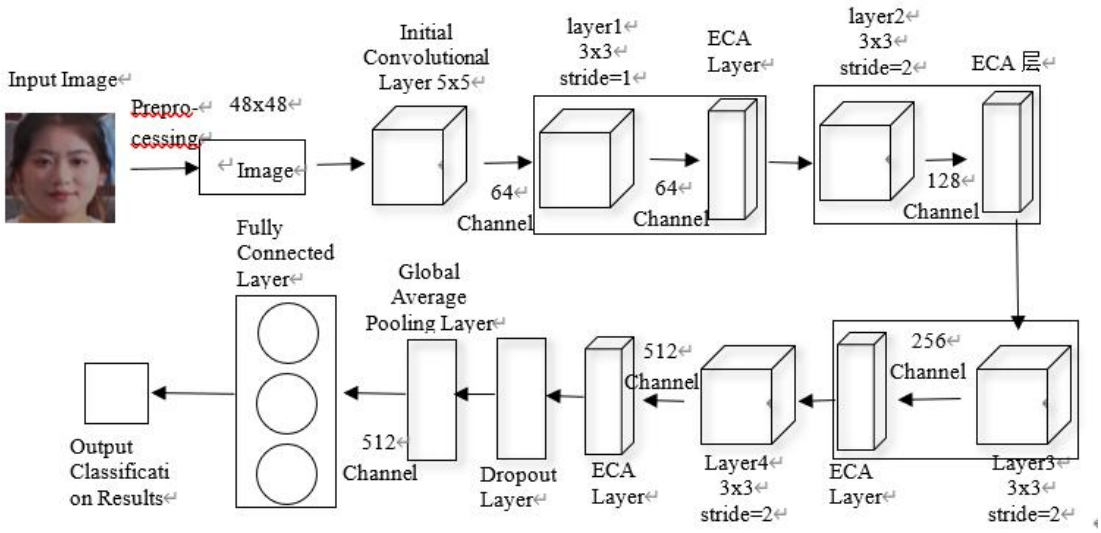


Figure 1 Network Flow Chart

The overall network flow is based on the ResNet18 architecture, including one initial 5x5 convolutional layer (64

channels), 4 residual stages (totaling 8 residual blocks, each with two 3x3 convolutional layers, channel numbers sequentially 64, 128, 256, 512), one global average pooling layer, and one fully connected layer. The input is a 48x48 single-channel facial expression image. After initial convolution, batch normalization and Mish activation are applied; each residual block embeds the ECA mechanism to enhance feature expression, followed by batch normalization and Mish activation; the shortcut uses 1x1 convolution for dimension adjustment. Finally, classification into 3 emotion categories is achieved through global average pooling, Dropout (probability 0.5), and the fully connected layer, significantly enhancing the accuracy and generalization ability of micro-expression recognition.

The ECA (Efficient Channel Attention) module is an optimized version of the SE (Squeeze-and-Excitation) module, designed to address the trade-off between performance and complexity in the SE module[12].

The SE Block is not a complete network but a substructure that can be embedded into other classification or detection models. The core idea of the SE Block is to allow the network to learn feature weights based on the loss, so that effective feature maps have larger weights, while ineffective or less effective feature maps have smaller weights, thereby training the model to achieve better results. The structure diagram is shown in Figure 2.

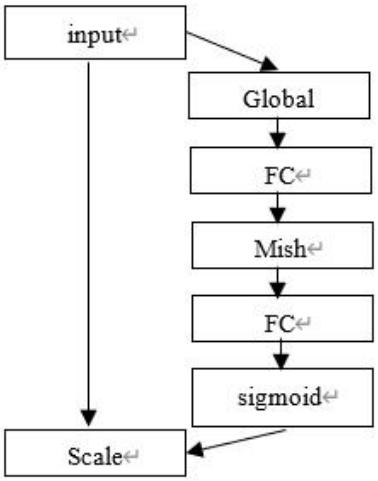


Figure 2 Structure of Improved Attention Module

The ECA module, as an optimized version of the SE module, significantly reduces computational complexity and parameter count through one-dimensional convolution and adaptive kernel size design, while retaining the core advantages of channel attention. Compared to the SE module, the ECA module avoids dimensionality reduction operations and effectively captures cross-channel correlations through a local cross-channel interaction strategy, reducing parameters by about 80% (e.g., ECA requires only 80 parameters compared to 24.37M for SE on ResNet50), and computation from 3.86 GFLOPs to 4.7e-4 GFLOPs, while increasing Top-1 accuracy by over 2%[13]. Experiments show that ECA outperforms SE and CBAM on the ImageNet classification task, with lower parameters and competitive performance compared to AA-Net. In object detection and instance segmentation tasks on the COCO dataset, ECA-Net, with ResNet50 and MobileNetV2 as backbones, demonstrates lower computational complexity and higher detection accuracy. Its adaptive kernel size enhances the model's adaptability to different channel numbers, facilitates integration into various CNN architectures, and reduces overfitting risk, making it particularly suitable for resource-constrained scenarios.

2.2 Training Process Improvements

2.2.1 Label smoothing loss function

During model training, a series of optimization strategies were employed to enhance the robustness of micro-expression recognition and alleviate overfitting on small datasets. Firstly, to address the issue of reduced model generalization capability caused by the excessive pursuit of extreme predicted probabilities in traditional cross-entropy loss functions for classification tasks, the Label Smoothing Loss technique was introduced. This technique softens the distribution of true labels by introducing a small amount of uniformly distributed noise into the training labels, thereby preventing the model from becoming overconfident about training samples. Specifically, the label smoothing loss function replaces the original hard labels with soft labels, i.e., it retains the dominant weight for the true class while assigning tiny probabilities to other classes. This encourages the model to learn more generalizable feature representations and reduces sensitivity to noisy samples.

2.2.2 Hybrid optimization strategy

To further accelerate model convergence and improve training efficiency, this system adopts a hybrid optimization strategy combining the AdamW optimizer and the OneCycleLR learning rate scheduler to achieve super-convergence. The AdamW optimizer, an improved version of the Adam optimizer, effectively mitigates the training instability that can arise from L2 regularization in traditional Adam optimizers by decoupling weight decay, thereby maintaining model generalization ability while enhancing training efficiency.

The OneCycleLR learning rate scheduler follows the principle of super-convergence. Within a training cycle, it first linearly increases the learning rate from a low value to a maximum value, and then gradually decreases it to a very small value, forming a complete learning rate cycle. This strategy allows for rapid exploration of the parameter space in the early stages of training, promoting fast model convergence, while fine-tuning parameters with a lower learning rate in the later stages helps escape local optima and improves the model's generalization ability.

By combining the parameter optimization advantages of the AdamW optimizer with the dynamic learning rate adjustment mechanism of the OneCycleLR scheduler, the hybrid optimization strategy constructed in this system fully leverages the synergistic effects of both. During model training, the AdamW optimizer adaptively adjusts the parameter update step size, effectively handling gradient sparsity issues for different parameters, while the OneCycleLR scheduler provides a more reasonable learning rate trajectory for the entire training process, enabling the model to achieve higher accuracy in fewer training cycles, ultimately achieving super-convergence. This hybrid optimization strategy not only significantly improves model training efficiency but also demonstrates stronger generalization ability and robustness in the emotion classification task.

2.2.3 Early stopping mechanism

Simultaneously, to prevent overfitting during the training process, this system incorporates an Early Stopping mechanism, which dynamically terminates training by continuously monitoring the validation set loss function. If the validation loss does not show significant improvement over a preset number of consecutive epochs, the best model parameters are automatically saved and the training process is terminated. This mechanism further optimizes the model's generalization performance on small datasets and enhances model stability and reliability.

3 EXPERIMENTS AND RESULT ANALYSIS

3.1 Dataset

Since expressions in the classroom are mostly discrete expressions, based on the valence-arousal and discrete expression correspondence model proposed by He Jiabei et al., shown in Figure 3, the dataset expressions were categorized into three types: positive, neutral, and negative. Several basic expression types were reclassified into these three categories according to their valence relationship.

The dataset was obtained through screenshots from online real-time class videos, part of the micro-macro expression MMEW dataset jointly released by Professor Ben Xianye's team from Shandong University, and photos taken by multiple volunteer students, manually annotated. It contains 120 images each for active, neutral, and negative expressions, totaling 1800 grayscale expression images after data augmentation. The dataset was split into training and test sets in an 8:2 ratio, and the ten-fold cross-validation method was used to partition the data multiple times and average the results, reducing bias introduced by randomness.

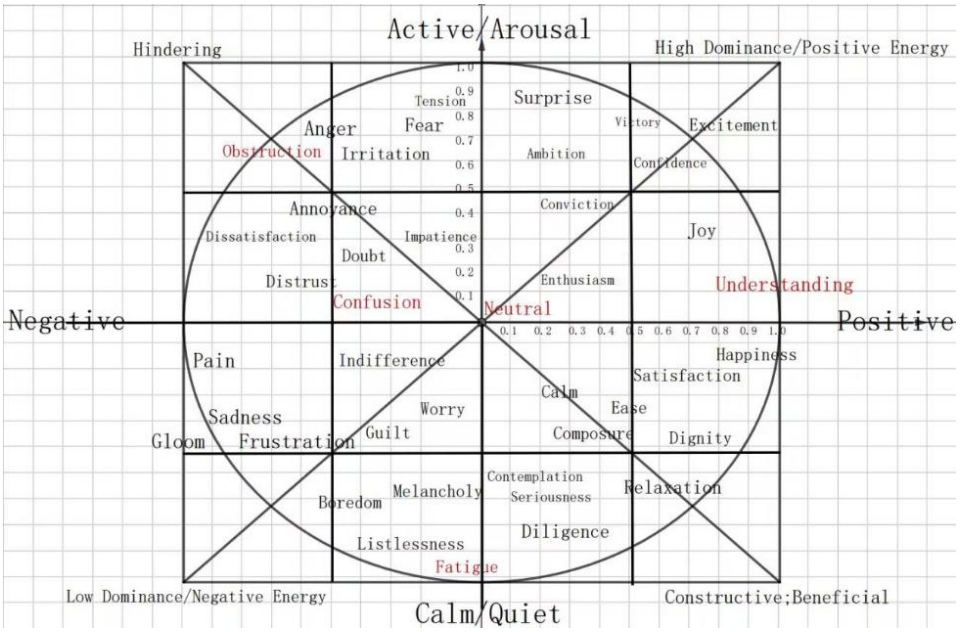
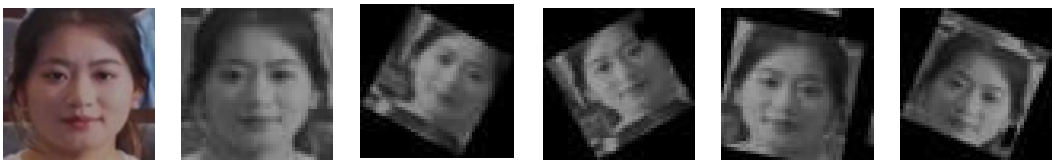


Figure 3 Correspondence Between Valence-Arousal and Discrete Expressions of He Jiabei et al.

3.2 Dataset Preprocessing and Training

Dataset preprocessing in this experiment mainly included: image standardization (grayscale conversion, normalization), multi-dimensional data augmentation (random horizontal flipping, random translation, random rotation, random erasing), then scaling to 48x48, and converting to Tensor format for input into the network for training. Data processing comparisons are shown in Figure 4, applying randomly defined data augmentation functions:



(a) (b) (c) (d) (e) (f)

Figure 4 Data Processing Comparison

Due to the small dataset size, ten-fold cross-validation was used for training and testing. The initial learning rate was 0.0001, serving as the starting point for the OneCycleLR scheduler to prevent gradient explosion in early training. The weight decay coefficient for the AdamW optimizer was 0.0001, the batch size was 16, and the early stopping patience was 12 epochs. The model was built based on the PyTorch framework, and all experimental results were obtained on an NVIDIA GeForce RTX 4060 GPU with 16GB VRAM.

3.3 Ablation Experiment

As shown by the confusion matrix, this experiment verified that through model optimization and selecting better hybrid optimization strategies, excellent results were achieved on the self-built small dataset, reaching an accuracy of 97.50%. Table 1 lists the experimental results for different combinations of sub-modules. The baseline ResNet18 model adapted for small-scale datasets by reducing convolution kernel size, adjusting downsampling methods, and adding Dropout is labeled as ResNet18(Dropout). The final version with Mish function and added ECA channel attention mechanism is labeled as ResNet18(ECA).

Table 1 Experimental Results of Different Combinations

Experiment Group	ResNet18(Dropout)	ResNet18(ECA)	Accuracy (%)
1	√		97.22
2	√	√	97.50

It can be observed that although ResNet18_ECA has about 519,000 fewer parameters than the standard ResNet18 (a reduction of approximately 3.6%), the accuracy does not decrease but instead increases by 0.28%. Although the improvement is limited, possibly due to suboptimal hyperparameter tuning and the small dataset size not fully leveraging the optimized module's performance, it still validates the effectiveness of the module.

Figure 5 shows the confusion matrix for this experiment. It can be seen that the neutral state bears the primary representation of the learning state. This is mainly because negative expression states encompass many features (e.g., yawning, sadness, disgust, anger), with large differences between these features. Furthermore, some expressions, like disgust (which might only differ slightly from neutral by the mouth corner) or yawning (similar to surprise with an open mouth), lead to slightly lower accuracy for the negative class, making it more challenging.

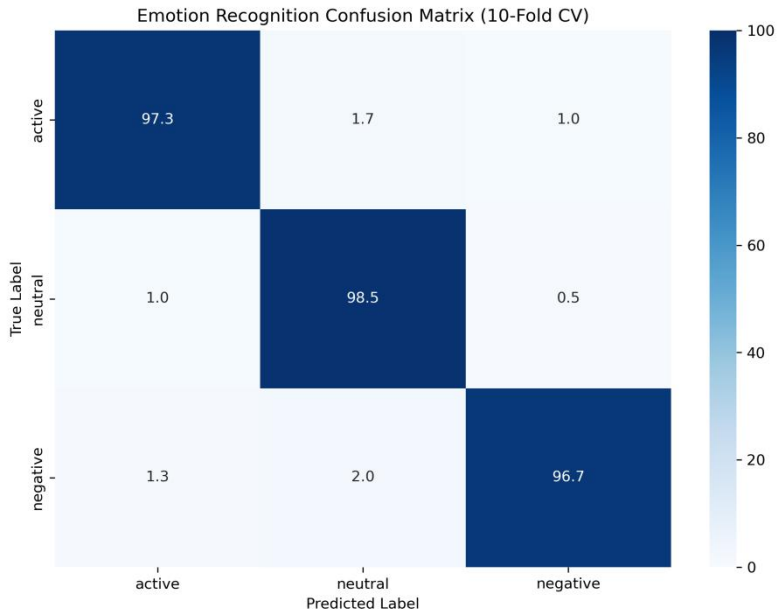


Figure 5 Confusion Matrix

4 CONCLUSION

This paper designed a method for recognizing learning and psychological states based on micro-expressions using a Residual Network. It uses single-channel input to extract facial information related to emotions and classifies them through valence relationships. Experimental results demonstrate that the proposed method exhibits strong generalization capability and applicability when processing large volumes of real-world facial expression data. The modified ResNet18 convolutional neural network algorithm, combined with the channel attention mechanism for automatic feature extraction, has achieved highly effective performance.Compared with traditional methods, the approach proposed in this study allows for more in-depth analysis of student engagement in class, providing teachers with better references for understanding student learning quality and formulating more effective teaching improvement strategies.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This article is an achievement of the Hubei Provincial First-Class Undergraduate Major Construction (Journalism), and a research outcome of the 2024 National College Student Innovation and Entrepreneurship Training Program Project “Detection of College Students’ Learning and Psychological States Based on Micro-expression Recognition” (Project No. 202412309008).

REFERENCES

[1] Tonguc G, Ozkara BO. Automatic recognition of student emotions from facial expressions during a lecture. Computers & Education, 2020(148): 1-12.
[2] Wei YT, Lei F, Hu MJ, et al. Review of Research on Student Expression Recognition. The Chinese Journal of ICT in Education, 2020(21): 48-55.
[3] Ekman P. Facial expression and emotion. American Psychologist, 1993, 48: 384.
[4] Tong XY, Sun SL, Fu MX. Data augmentation and second-order pooling for facial expression recognition. IEEE Access, 2019, 7: 86821-86828.
[5] Huang XL, Gou XS, Chen X. Facial Micro-expression Recognition Algorithm Based on Hybrid Features and Information Entropy. Computer Simulation, 2023, 40(06): 197-201.
[6] Qiao GF, Hou SM, Liu YY. Facial expression recognition algorithm based on improved convolutional neural network and support vector machine. Journal of Computer Applications, 2022, 42(04): 1253-1259.
[7] Chen T, Xing S, Yang WW, et al. Facial expression recognition integrating spatiotemporal features. Journal of Image and Graphics, 2022, 27(07): 2185-2198.
[8] Yang HY, Ciftci U, Yin LJ. Facial Expression Recognition by De-expression Residue Learning. Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 2168-2177.
[9] He KM, Zhang XY, Ren SQ, et al. Deep Residual Learning for Image Recognition Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770-778.
[10] Wang QL, Wu BG, Zhu PF, et al. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, 2020: 11531-11539. DOI: 10.1109/CVPR42600.2020.01155.
[11] Misra D. Mish: A Self Regularized Non-monotonic Activation Function. 2019: 8. https://arxiv.org/abs/1908.08681.
[12] Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7132-7141.
[13] He JB, Zhou JX, Gan JH, et al. Classroom Expression Classification Model Based on Multi-task Learning. Journal of Applied Sciences, 2024, 42(06): 947-961.