ITC 3/54
Information Technology and Control
Vol. 54 / No. 3/ 2025
pp. 844-863
DOI 10.5755/j01.itc.54.3.41283

Dense-Attention CNN with Spatial-Attention Fusion for Robust Facial Expression Recognition

Received 2025/04/23 | Accepted after revision 2025/07/19

HOW TO CITE: Li, D., Sun, J., Liu, W., Wang, L., Zhou, N. (2025). Dense-Attention CNN with Spatial-Attention Fusion for Robust Facial Expression Recognition. *Information Technology and Control*, 54(3), 844-863. https://doi.org/10.5755/j01.itc.54.3.41283

Dense-Attention CNN with Spatial-Attention Fusion for Robust Facial Expression Recognition

Dan Li*, Jinping Sun*

School of Information Engineering (School of Big Data), Xuzhou University of Technology, Xuzhou, Jiangsu, China.

Weiwei Liu, Likai Wang, Nan Zhou

Traffic police detachment of Xuzhou Public Security Bureau, Xuzhou Jiangsu, China.

Corresponding author: lidanonline@xzit.edu.cn (Dan Li); sjp@xzit.edu.cn (Jinping Sun)

Abstract: Currently, facial expression recognition technology has been gradually applied in fields such as intelligent healthcare, online education, and assisted driving. However, traditional Convolutional Neural Network (CNN) lack attention to facial local regions related to emotions, and classic loss functions cannot handle intra-class variability in facial expressions. This paper establishes a facial expression recognition model combining deep learning and attention mechanisms for both static and dynamic facial expressions. By extracting image features, it obtains rich multi-scale information flow and controls the number of model parameters. It constructs a spatial attention unit to focus on information with significant emotional intensity, and combines an intra-class distance penalty term and classification loss to supervise the network learning process. This approach addresses the issue of CNN paying insufficient attention to regions of interest while reducing the variability among facial expressions of the same class. Experimental results show that the accuracy of this model has increased by 1.1% and 2.7% on the CK+ and FER2013 public datasets, respectively.

KEYWORDS: Expression recognition, Deep learning, Convolutional neural network, Attentional mechanism

1. Introduction

Facial expression is an important form of expression that reflects people's inner feelings and psychological activities [6]. The most natural and common way for human beings to convey emotion and intention is facial expression. Aouani's research shows that in people's daily communication, facial expressions convey emotions, accounting for the highest proportion [2]. Therefore, facial expressions play an irreplaceable role. In recent years, with the rapid development of machine learning, facial expression recognition technology has also made great progress and has broad application prospects. Compared with traditional feature extraction (such as LBP, HOG) and detection methods (such as SIFT, Haar), deep convolutional neural networks (DCNN) can automatically learn multi-level and abstract expression features, are more robust to pose and lighting changes, have higher classification accuracy, but also have higher computational complexity. Traditional methods rely on manually designed features and have weak adaptability. Deep learning has significantly promoted the development of facial expression recognition. By densely connecting convolution and spatial attention mechanism, the model can accurately capture local muscle deformation features such as eye and mouth circumference. Combined with intra class distance constraint loss function, it effectively reduces the variance of similar expression features. Integrating 3D convolution and spatiotemporal attention modules to achieve temporal feature modeling for dynamic scenes. The current research focuses on model lightweighting and cross-cultural adaptation, which is accelerating the implementation of this technology in fields such as smart transportation, medical diagnosis, and educational emotional computing.

Real-time emotion analysis in HCI, monitoring of driver safety status, and pain warning intervention in medical scenarios are driving the development of facial expression recognition technology. Through multimodal signal fusion and micro expression spatiotemporal modeling in medical monitoring, pain threshold breakthrough signals can be captured in advance, improving the timeliness of clinical intervention; An attention analysis model that integrates facial action units and head posture in educational settings achieves an accuracy of 89.2% in recognition of concentration in real classrooms; In the field of safe driving, real-time facial expressions are used to monitor and warn of fatigue or dangerous states; Human computer interaction devices can automatically adjust screen brightness or push emotion related content based on eye conditions; The smart retail system triggers product recommendations by capturing customers' positive expressions; Key security areas will achieve millisecond level analysis of crowd emotions within 10 meters, and emergency plans will be automatically activated with abnormal panic expressions. These applications indicate that building high-precision facial expression recognition models has become a key technology for parsing human behavioral intentions.

After the 21st century, with the rapid development of science and technology, facial expression recognition technology has also ushered in a wave of research upsurge [3, 28]. Zhao et al. [38] proposed a facial expression recognition method based on local binary patterns and kernel discriminative isometric mapping, which improves expression classification performance by integrating local texture features with nonlinear manifold learning. Krizhevsky et al. [19] constructed a deep convolutional neural network, which significantly improved recognition accuracy in ImageNet image classification tasks through multi-layer cascaded convolution and GPU accelerated training. Simonyan et al. [29] designed the VGG convolutional neural network architecture, using a small-sized convolution kernel stacking strategy, and verified the critical role of network depth in large-scale image recognition tasks. He et al. [16] proposed a residual learning framework to address the difficulty of training deep networks. Goodfellow et al. [15] analyzed data from three machine learning competitions, revealing key challenges such as adversarial samples and feature representation robustness, providing important benchmarks for defensive design of deep learning models. Tzirakis et al. [34] proposed an end-to-end Bi modal feature layer fusion method for voice and facial expression. Using CNN [37,13] and 50 layer ResNet network to extract emotional features, the Bi modal features are cascaded directly, and the two-layer LSTM model is used to model and train the system. Zhu [39] pro-

posed a new 3d-fer method, that is, a convolutional neural network based on differentiated attention (DA-CNN) to generate a more comprehensive expression correlation representation. However, due to the expansion of the dimension, parameters of the 3D convolutional neural network are typically excessively numerous, while the sample size of the existing expression video database is extremely limited, and there is a lack of relevant pre trained models, resulting in the over fitting phenomenon of directly training the 3D convolutional neural network model, and it is not sufficient to learn the temporal changes between image frames.

Currently, deep learning still faces multiple challenges in the field of facial expression recognition: 1) significant environmental interference, including complex lighting conditions, facial occlusions (such as masks), and non face poses; 2) The inherent defects of the dataset, such as differences in facial features and micro expression emotional representations across regions and ethnicities, exacerbate the difficulty of recognition; 3) At the algorithmic level, there is sensitivity to hyperparameters and potential algorithmic bias in comparative analysis methods, leading to reduced reproducibility of results. These factors collectively constitute the core bottleneck for the implementation of facial expression analysis technology. A series of predictable or uncontrollable problems have led to a significant gap between the accuracy of facial expression recognition and that of object recognition. For example, CNN has a maximum accuracy of 97% for the fashion MNIST object data set [10,30,31], but only 72% ± 1% for the FER2013 facial expression data set [4,20,26]. The accuracy difference of nearly 25% can clearly show that the research of facial expression recognition still has great potential and future.

This study proposes an expression recognition model based on lightweight dense connected convolutional neural networks. To address the problem of traditional CNN's insufficient attention to emotion related local regions and the difficulty of classical loss functions in handling intra class differences, multi-scale feature extraction is used to control parameter quantities and obtain rich information flows. A spatial attention mechanism is introduced to focus on emotion significant regions, and a joint optimization is carried out by combining intra class distance penalty term and classification loss. The experiment validated the model on a public facial expression dataset, which effectively improves the attention to key areas and reduces intra class expression differences while maintaining lightweight, achieving better recognition performance.

2. Facial Expression Recognition

2.1. Traditional Methods

Traditional facial expression recognition methods generally involve four steps: data preprocessing, feature extraction, feature selection, and training an appropriate classifier.

1 **Data preprocessing**

Firstly, image normalization is applied to unify the data format and reduce the impact of variations in illumination and resolution. Subsequently, denoising techniques are utilized to smooth the image and suppress noise. Finally, face detection (such as Haar Cascade classifier) is employed to locate the facial region, and keypoint localization (such as ASM/AAM model) combined with affine transformation is used to align the facial pose, ensuring that subsequent feature extraction is based on a consistent spatial structure.

2 **Feature extraction**

The core of feature extraction lies in mining key information related to expressions from preprocessed images. Geometric features directly quantify geometric deformations of expressions (such as widened eyes in surprise) by calculating distances, angles, or motion trajectories of facial keypoints (such as the corners of the eyes and mouth). Appearance features utilize LBP, HOG, or Gabor wavelets to describe local texture or edge changes, capturing surface patterns induced by muscle movements. Statistical features employ dimensionality reduction techniques such as PCA and LDA to project high-dimensional image data into a low-dimensional space, retaining the most discriminative global features (such as Eigenfaces).

3 **Feature selection**

Feature selection aims to filter out the most classification-valuable subset from high-dimensional

features, in order to reduce computational complexity and avoid overfitting. Filter-based methods (such as chi-square test, mutual information) directly evaluate the correlation between features and categories through statistical indicators, retaining highly correlated features; wrapper-based methods utilize classifier performance to iteratively optimize the feature subset, dynamically adjusting feature importance.

4 **Train classifier**

Models such as SVM, k-NN, Random Forest, or AdaBoost are commonly used. These classifiers predict new samples by learning the association rules between features and expression labels.

Traditional facial expression recognition methods are highly interpretable, require low computational resources, and exhibit robustness to small sample data. However, deep learning methods, through end-to-end automatic feature learning, overcome the limitations of manually designed features, capture more complex nonlinear relationships, and significantly enhance recognition accuracy and generalization ability in big data scenarios.

2.2. Deep Learning Methods

Deep learning originates from neural networks, which learn high-level abstract features through multi-layer structures, and CNN [12, 21, 25], as a part of deep learning, performs particularly well in image processing.

1 **Convolutional Neural Network**

In addition to the input layer, convolution layer, pooling layer, fully connected layer, and output layer, CNN [24, 7, 18] also introduces activation function [8, 22, 36] during the convolution calculation process to increase the nonlinearity of the neural network model. The CNN structure includes the following parts.

1) Input layer: Perform targeted data processing on input data from different dimensions. In convolutional neural networks, input features should be uniformly normalized and preprocessed. 2) Convolutional layer: Each convolutional layer in a convolutional neural network is composed of several convolutional units. The convolution operation [32, 17] uses convolution kernels to multiply regions of the same size in the corresponding feature map, and adds the resulting values to obtain the input corresponding to the output feature map. 3) Pooling layer: The pooling layer is generally located between two adjacent convolutional layers, which can effectively reduce the size of the parameter matrix and model, improve robustness, and reduce overfitting. 4) Fully connected layer: Each node is connected to all nodes in the previous layer to synthesize the features extracted earlier. The main function of full connectivity is to improve network classification performance and complete classification tasks. 5) Output layer: The output layer of a convolutional neural network focuses on different data features for different input data [35]. If the input is image data, a normalization function such as softmax is usually used for image feature classification, and the results of these multi classifications are displayed in probability form.

The function of the activation function is to incorporate nonlinear factors, allowing the convolutional neural network to fit any nonlinear function. The common activation functions include sigmoid function and ReLU function [11, 33]. The mathematical expression of the sigmoid function is shown in Formula (1), which converts the input value into an output value in the 0-1 range. The ReLU function is also a common nonlinear activation function in convolutional neural networks, and its mathematical expression is based on Formula (2).

$$S(x) = \frac{1}{1+e^{-x}} \tag{1}$$

$$f(x) = \max(0, x) \tag{2}$$

The training of CNN relies on the BP algorithm. Assuming that forward propagation does not obtain the expected output value, the square error between the expected output value and the current output value of the neural network can be calculated, which is the loss function [27, 5]. The calculation formula refers to Formula (3).

$$E = \frac{1}{2n}\sum_{x} y(x) - h_{w,b}{}^{2}, \tag{3}$$

where E is the loss function, and x represents the input of the network. The weight matrix and the offset matrix are represented by w and b respectively. $h_{w,b}$ indicates output. The target output value is $y(x)$. n is the number of samples.

The backpropagation algorithm calculates the partial derivatives of the objective function with respect to the weights of each neuron layer by layer, forming the gradient of the objective function with respect to the weight vector, which serves as the basis for modifying the weights. The learning of the network is completed during the weight modification process. Calculate the partial derivatives of the weight matrix and bias matrix for each neuron separately. As shown in Formulas (4)-(5).

$$w^l_{ij} = w^l_{ij} - \alpha \frac{\partial E}{\partial w^l_{ij}} \tag{4}$$

$$b^l_i = b^l_i - \alpha \frac{\partial E}{\partial b^l_i}, \tag{5}$$

where w^l_{ij} represents the weight from the j-th neuron of layer l to the i-th neuron of layer (l+1), and b^l_i represents the offset value from the j-th neuron of layer a to each neuron of layer l. α is the learning rate.

The partial derivatives in Equations (4)-(5) are the partial derivatives of the weights and bias values, respectively. According to the above formula, calculate the partial derivatives of each layer's objective function with respect to the weight and bias parameters, and continuously update the parameters to minimize the loss function.

2 **Deep learning**

a **Data preprocessing and enhancement**

Data preprocessing is the foundation for improving model performance, eliminating lighting and scale differences through standardization (Z-score) or normalization (Min Max); Combining face detection (MTCNN) and keypoint detection (Dlib) to locate and correct facial poses; Finally, data diversity is expanded through random cropping, flipping, color perturbation, and GAN synthesis of samples to alleviate the problem of overfitting in small samples.

b **Model Architecture and Feature Extraction**

Deep learning automatically extracts hierarchical features through neural networks: CNN (such as VGG, ResNet) captures local textures (eye wrinkles), ResNet residual connections alleviate gradient vanishing; Lightweight design (MobileNet) adapted for mobile devices; Transformers (such as Swin Transformer) model global dependencies through self attention; RNN/LSTM combined with attention mechanism captures inter frame dynamics (such as quickly opening eyes when surprised).

c **Training Strategy and Optimization**

Training optimization is achieved through loss functions, regularization, and transfer learning: based on cross entropy, combined with focus loss to alleviate class imbalance; Dropout prevents overfitting, BatchNorm accelerates convergence; Use ImageNet pre trained models (such as ResNet-50) to initialize weights, freeze the underlying convolutional layers during fine-tuning, and only train the top-level classifier to improve performance on small datasets.

d **Post processing and ensemble learning**

Post processing improves robustness through model fusion and attention mechanisms: integrating CNN, Transformer, and RNN prediction results (such as voting or weighted averaging), utilizing the advantages of multiple architectures; The CBAM attention mechanism focuses on key areas (upward corner of the mouth); Class Activation Graph (CAM) visualization model focuses on the region of interest, assisting in the interpretation of medical/judicial scenarios.

3. Attentional Mechanism

When training and designing different convolutional networks for different tasks [14, 23, 9], researchers will adopt feature enhancement methods on the basis of the original network to enhance network performance and improve accuracy. Attention mechanism [1] is a neural network feature enhancement method.

3.1. Spatial Attention Mechanism

The fully connected spatial attention mechanism acts at the end of the convolutional neural network. Define the network output eigenvector as x, and the general calculation formula of the fully connected attention mechanism mask is as Formula (6):

$$a = g(W_a x + b_a), \tag{6}$$

where, W_a and b_a respectively represent the attention parameters, and $g(\cdot)$ represents the nonlinear function, which adjusts the attention weight to a certain range. The attention vector a has the same dimension as x, and the weighting process is calculated as Formula (7):

$$\tilde{x} = a \cdot x \tag{7}$$

Convolution spatial attention mechanism can act on any layer of convolution neural network. The characteristic graph of network layer output is defined as $X=[x^1, \ldots, x^C]$, and the general calculation formula of convolution attention mechanism mask is as Formula (8):

$$A = g(W_a * X), \tag{8}$$

where W_a represents the attention parameter and * represents the convolution operation with the same filling to keep the output mask consistent with the spatial dimension of the feature. By compressing the channel information, the attention plane A matches the spatial information of X, and the weighting process is calculated as Formula (9):

$$\tilde{X} = [\tilde{x}^1, \ldots, \tilde{x}^C] = [A \otimes x^1, \ldots, A \otimes x^c], \tag{9}$$

where, $\otimes$ represents the multiplication of corresponding elements. In comparison to the fully connected spatial attention mechanism, the convolutional spatial attention mechanism preserves the spatial structural information inherent in the original input image, thereby facilitating more effective learning of attention weights.

3.2. Spatial Attention Mechanism

Channel attention mechanism is a technique used in deep learning to enhance the feature expression ability of models. The core idea is to dynamically adjust the importance of different channel features, so that the model focuses on the more critical channels for the task, while suppressing redundant features. Its typical implementation is the Squeeze and Excitation (SE) module, which is widely used in CNNs such as ResNet and MobileNet. CNN features from different channels have different extraction effects, and the principle of using them is to select different semantic attributes for different content to achieve better results.

Using Global Average Pooling (GAP) as the main component of channel attention to generate channel descriptors $z \in R^c$. Among them, $u_c \in R^{H \times W}$ is the feature map of the c-th channel, and z_c represents the global average value of that channel as Formula (10).

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j). \tag{10}$$

Learn the weights s of each channel through a two-layer fully connected network (Botleneck structure), where each element $s_c \in [0,1]$ of s represents the importance of the c-th channel.

$$s = F_{ex}(z, W) = \sigma(W_2 \cdot \delta(W_1 \cdot z)). \tag{11}$$

In Formula (11), $W_1 \in R^{\frac{C}{r} \times C}$ represents the fully connected weight of the first layer, and r is the compression ratio (such as r=16), used for dimensionality reduction to reduce computational complexity. δ is the ReLU activation function, adding nonlinearity. $W_2 \in R^{C \times \frac{C}{r}}$ is the fully connected weight of the second layer, restoring the channel dimension. σ is the Sigmoid activation function, which normalizes the weights to [0,1].

Multiply the learned channel weights s with the original feature map u to enhance key channels and suppress redundant channels as Formula (12).

$$\tilde{X}_c = F_{scale}(u_c, s_c) = s_c \cdot u_c. \tag{12}$$

Final output feature map $\tilde{X} = [\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_c]$.

3.3. Self attention

Self attention mechanism is a technique used in deep learning to dynamically build global dependencies between features. By calculating the cor-

relation between different positions (or channels) in the feature map, attention weights are automatically assigned to focus the model on the more critical information for the task. Its core lies in "self-learning" feature association, without the need for manual prior design.

Assuming the input feature is $X \in R^{C\times H\times W}$, the self attention mechanism is implemented through the following steps:

1 **Linear transformation**

Map features into three vectors: Query (Q), Key (K), and Value (V) :

$$Q = W_Q X, K = W_K X, V = W_V X \tag{13}$$

In Formula (13), W_Q, W_K, and W_V are learnable projection matrices.

2 **Similarity calculation**

Calculate the correlation (attention weight) between Query and Key through dot product or scaled dot product:

$$A = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right). \tag{14}$$

In Formula (14), d_k is the dimension of the Key used to stabilize gradients.

3 **Feature weighting**

Multiply attention weight A with Value to generate weighted features :

$$Z = AV. \tag{15}$$

In Formula (15), the final output Z integrates global dependency information.

4. Design of Facial Expression Recognition Model

4.1. Overall Design

The process of facial expression recognition includes data set preprocessing, model training and verification, and facial expression recognition.

Data set loading and preprocessing module. It is necessary to complete the work of loading and processing data set, convert the original data into data information more suitable for neural network, reduce over fitting and speed up operation. At the same time, the expression is classified (for example, the FER2013 data set divides the expression into angry, distinct, scaled, happy, sad, surprised, neutral). It is worth mentioning that different datasets may have different facial expressions.

Model training and verification module. To train a neural network model, it is essential to load the network structure, import the training dataset, construct the model along with its optimizer, and proceed with the training process. Subsequently, the test dataset is employed to evaluate the accuracy of the trained neural network model. The model with higher accuracy than the previous training will be saved for the next facial expression recognition. At the same time, TensorBoard, a visual tool, is used to monitor what happens inside the model in a visual way during the training process.

Facial expression recognition module. Output the local face pictures or the face information recognized by the camera to the facial expression recognition module, load the trained recognition model, recognize and display the facial expression, and the system can display the percentage of each expression in the total number of people.

The following is the overall structure design of the model, as shown in Figure 1.

4.2. Expression Database

The experiment mainly uses CK+, FER2013 and self built data sets to train the network and analyze the experimental results. The CK+dataset was chosen because it contains finely annotated dynamic expression sequences, which are suitable for capturing micro expression changes; FER2013 can verify the generalization of the model through large-scale, multi scene static images (including pose/lighting changes). Both are authoritative benchmark datasets in the academic community, widely used for result reproduction and horizontal comparison. Their combined use can balance model accuracy and robustness verification.

The CK+ dataset was expanded on the basis of Cohn Kanade dataset and released in 2010. The dataset

Figure 1
Overall structure.

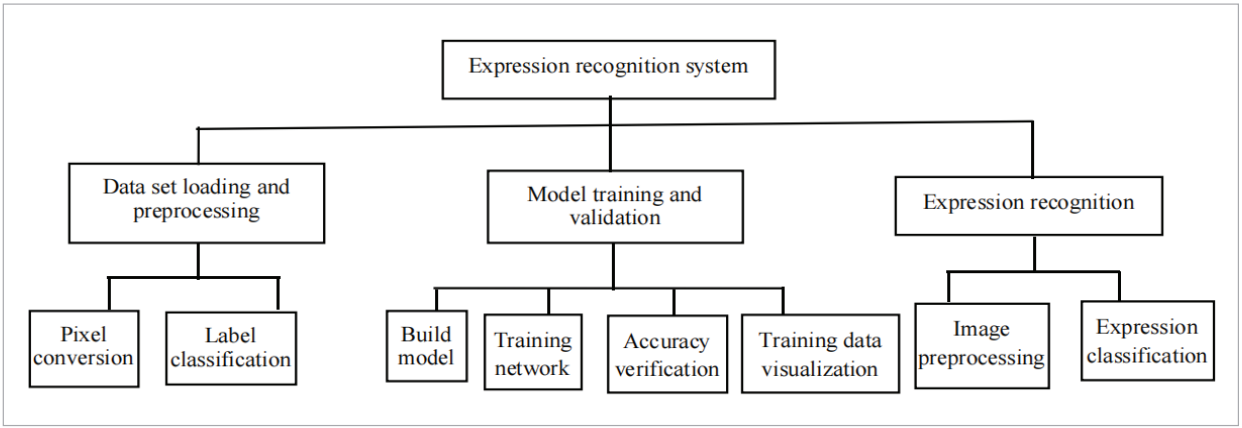


Figure 2
Partial images of CK+ dataset.

contains 123 objects and 593 image sequences. The CK+ data set comes from 123 subjects, mainly from western countries, and is divided into 7 expressions, including surprise, contempt, anger, fear, happiness, disgust and sadness. A total of 327 facial expression sequences were marked. In each sequence, researchers recorded the process of subjects' gradual transformation from normal expression, that is, neutral expression, to a certain required expression, and took the last image of each sequence as the data picture of the required expression. This data set is a popular data set in the research of facial expression recognition. Figure 2 shows some facial expression images in the CK+ dataset. Additionally, since the picture information is represented in a numerical matrix format of type uint8, with values ranging from 0 to 255, and data processing typically employs the double type, it is necessary to divide the pixel values by 255 in order to accurately represent the picture information in the desired format.

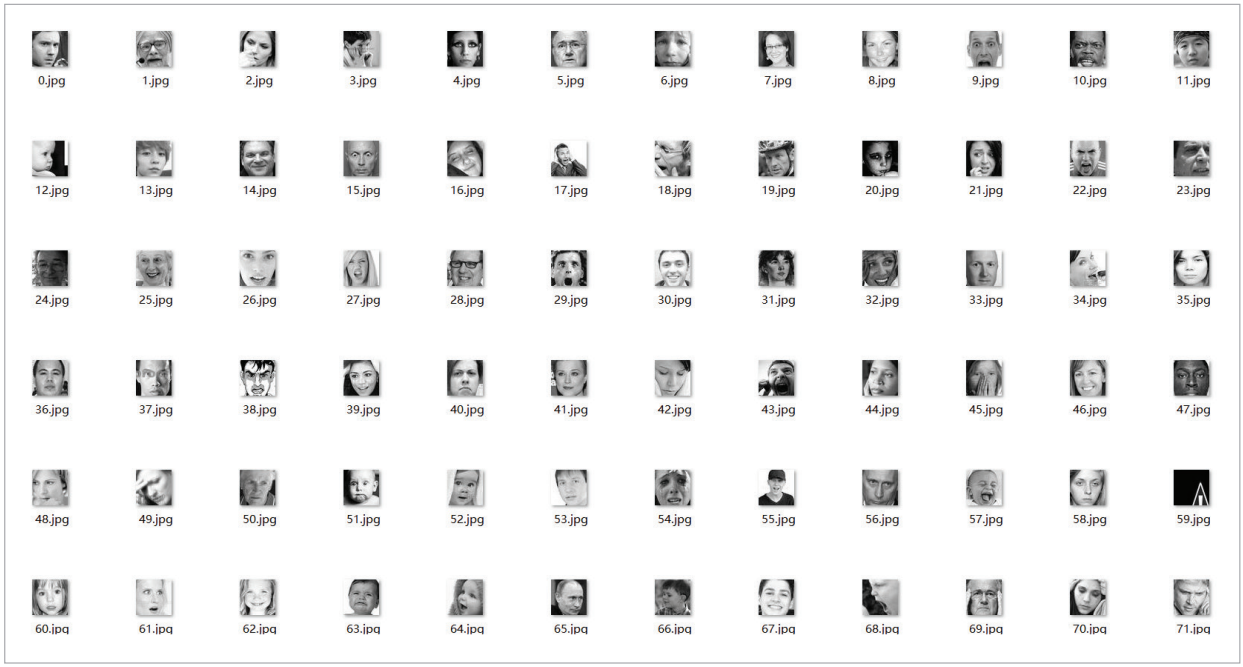
The FER2013 dataset is the official dataset offered by the 2013 Kaggle competition focused on facial expression recognition. Most of the pictures are obtained from web resources through crawlers, so it contains facial expressions of different races, ages and angles, and there are some interfering pictures, including non real people, large-area occlusion, non-human faces, etc., so the experimental error is large. The FER2013 dataset contains more than 35000 images and is divided into three parts, including 28708 Training sets, 35898 PublicTest sets and 35898 PrivateTest sets. Every image undergoes preprocessing to be transformed into a grayscale image with a uniform dimension of 48*48. The FER2013 dataset is classified into 7 kinds of expressions, which are respectively represented by digital labels 0-6, they are: 0 anger; 1 dislike; 2 fear; 3 happy; 4 sad; 5 surprised; 6 neutral. FER2013 facial expression dataset is a CSV file provided by Kaggle, in which each line contains the expression classification label, pixel information and classification purpose of the dataset.

Use pandas to store the expression classification and picture pixel information in label CSV and data CSV. The pixel information is transformed into pictures visible to human eyes and classified according to the training set and test set, as shown in Figure 3.

4.3. Design of Expression Recognition Model Based on Lightweight Dense Connected Convolutional Neural Network

The convolutional network trained for facial expression recognition typically demonstrates a robust capability to fit data effectively. Nonetheless, owing to constraints imposed by the optimization algorithm and computational resources, the computational power often becomes a limiting factor for the model when tackling intricate tasks. To address this, the convolutional structure and pooling layers with local connectivity inherent in convolutional networks can be leveraged to streamline the network architecture, thereby mitigating the tension between model complexity and expressive power. For tasks involving network models, this approach not only allows for an increase in model complexity but also enhances the model's expressive capabilities. The local to global multi-level integrated features of facial appearance are obtained through the dense connection convolution module, and the matching spatial attention units are constructed to focus on the high emotion related

Figure 3
Partial images of FER2013 dataset.

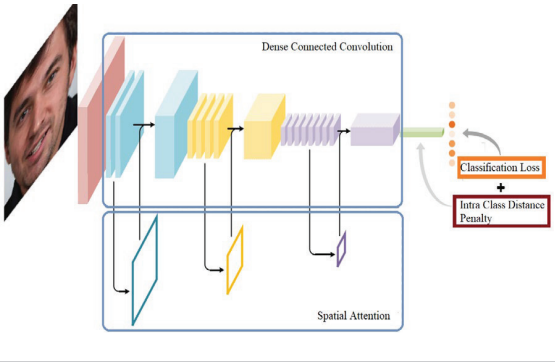


areas. In addition, the performance of the model is further improved by designing a joint loss function to reduce the variation within the feature class.

The expression recognition model architecture includes three parts: spatial attention unit, dense connection convolution module and output layer. The spatial attention unit tracks the local area of interest related to the student's facial expression without increasing the number of parameters. The convolution module is densely connected to obtain the multi-scale information of the input image. The output layer can minimize the changes within the class, so as to transmit the facial features of the same class to the class center. In addition, the classification error is minimized, and the expression polarity probability distribution is generated. Introducing spatial attention units can focus on emotionally significant areas (such as around the eyes and mouth), enhance key feature expression through weighted reinforcement, suppress irrelevant noise, and improve feature discriminability; The intra class distance penalty term constrains the convergence of features from similar samples towards the class center, reducing intra class variance caused by individual differences (such as facial structure) and enhancing the model's generalization ability to essential facial features.

The model architecture is shown in Figure 4, which includes three core components: firstly, the dense convolution module adopts a layer by layer fusion of multi-level convolution features to achieve multi-scale feature extraction of the input image; Secondly, the spatial attention mechanism dynamically locates expression related regions (RoI) while maintaining lightweight parameters through interactive modeling of spatial positional relationships, enhancing facial emotion feature extraction; Finally, the output module optimizes the distribution of intra class features to cluster similar samples towards the class center, and combines it with the classification loss function to generate facial expression polarity prediction results. To standardize input specifications, all images are scaled to a resolution of 224 × 224.

Figure 4
System model architecture.



1 **Dense connection convolution module**

In CNN, convolution layer and pooling layer are superimposed alternately. The information of low-level feature map is concrete but semantic is fuzzy. The high-level feature map is abstract and semantic, but insensitive to local dependencies. By introducing the dense connection structure, we can make better use of the low-level features. The dense connection mode of dense connection blocks enables the features in the network to be fully spread and used. Compared with the features that the general depth network directly depends on the output of the last layer, each convolution layer in the dense connection layer only needs to add fewer convolution cores in turn to obtain a larger information flow, and the number of parameters is greatly reduced. The dense connection network can be improved by using more low-level features, and the trained classifier has more generalized classification performance. In the dense connection layer, the current convolution layer input is the feature graph set output from all previous convolution layers, and the output with a fixed number of channels is generated after convolution as part of the input of the next convolution layer. The output of the ith convolution layer in the jth dense layer can be expressed as Formula (16):

$$H_j^i = Conv(Concat(H_j^0, H_j^1, ...H_j^{i-1})) \quad (16)$$

where $Conv(\cdot)$ is a composite function, which represents all operations in the convolution layer. $Concat(\cdot)$ represents the merging of characteristic graphs in the channel dimension, and H_j^0 represents the input of this layer.

Dense connected networks will not have the problem of smaller and smaller gradients with the in-

crease of network layers. The detailed information near the input end is transferred to a deeper level to complement the global information, which can realize feature reuse and enhance the dissemination of information.

The transition layer divides the structure of the dense connection network, including batch normalization, linear rectification function, convolution layer and pooling layer. In order to improve the compactness of the model and reduce the number of feature maps, the convolution layer uses 1*1 convolution kernel to further refine the information in the feature map, and sets a dense connection network to output n feature maps. The number of output characteristic maps generated after 1*1 convolution operation shall not exceed the maximum integer of θn, $\theta(0<\theta\leq 1)$ is the compression factor. The count of convolution kernels aligns with the number of channels in the input feature map, a configuration that does not compromise feature representation. Meanwhile, the pooling layer serves to diminish the spatial dimensions of the feature map, yielding scale-invariant features, accelerating information flow throughout the network, and easing the complexity of network training.

2 Spatial attention unit

The contribution of each feature to the final classification is different. In terms of separability, the features with high separability contribute more than other features. The principle of spatial attention mechanism is a signal processing mechanism that simulates the human visual system (HVS) and pays more attention to the visual areas related to task targets. Because facial expression is composed of specific muscle movements, the features extracted from local muscle regions can best represent the expression information. By quantifying the importance of different spatial positions in the feature map, we can focus on the local area of the face with rich emotional information, which is conducive to the recognition task.

The spatial attention mechanism demonstrates significant technological advantages in facial expression recognition tasks by simulating the selective attention characteristics of the human visual system. The core innovation of this mechanism lies in the construction of a dual branch collaborative architecture, in which the main branch adopts a bottleneck convolution structure of 1 × 1-3 × 3-1 × 1, achieving channel dimension compression and expansion while maintaining the spatial resolution of the feature map. This design enables the network to efficiently capture local texture changes caused by facial expressions and movements, such as subtle contractions of the orbicularis oculi muscle and stretching deformations of the zygomaticus major muscle. The mask branch aggregates information from the channel dimension through parallelized global average pooling and global maximum pooling operations, generating a spatial attention prior map. The dual path architecture achieves decoupling between feature extraction and attention generation, retaining complete spatial contextual information while highlighting discriminative feature channels for facial expression classification.

$G_j=[H_j^O,\ldots,H_j^n]$ is defined as the output feature map generated by the jth dense layer. Firstly, the channel information is cooperatively compressed at all spatial positions of the characteristic map of different scales. The calculation process is as Formula (17):

$$M^i=Concat(AvgChannel(H_j^i),\\ MaxChannel(H_j^i)),i=0,1,\ldots n \tag{17}$$

$AvgChannel(H_j^i)$ and $MaxChannel(H_j^i)$ represent channel based global average pooling and global maximum pooling operations. Generate an attention map S^i for each M^i as Formulas (18)-(19).

$$S^i=DilatedConv(W,M^i) \tag{18}$$

$$S_{h,w,d}^i=\sum_{k,r,c}^{K-1,R-1,C_m-1}W_{k,r,c,d}\cdot\\ \cdot M^i_{h+k\varepsilon-\frac{(K+1)\varepsilon}{2},w+r\varepsilon-\frac{(R+1)\varepsilon}{2},C},d=1,\ldots D \tag{19}$$

where $DilateConv(\cdot)$ represents the hole convolution operation, and the region of interest is determined by using the spatial context, W represents the $k\times R$ convolution kernel, and D represents the number of hole convolution cores, that is, the number of channels of S^i. C^i indicates the number

of M^i channels. By using hole convolution, we can control the parameters and expand the receptive field at the same time, and effectively explore the dependence of longer distance spatial information. The attention plane A only related to the spatial position is formed through S^i, which is calculated as Formula (20):

$$S = Concat(S^0, ..., S^i ..., S^n) \tag{20}$$

$$A = PointwiseConv(W^p, S) \tag{21}$$

$$A_{h,w} = \sum_{c}^{C_s - 1} W_c^p S_{h,w,c} \ . \tag{22}$$

In Formula (20), $PointwiseConv(\cdot)$ represents the convolution operation of 1×1, mapping multiple channels of the attention graph to a single channel, W^p represents the convolution kernel, and C_s represents the number of channels of S. Finally, the attention weight is converted to the (0,1) interval to obtain the attention mask, which is calculated as follows:

$$m = \sigma(A) \tag{23}$$

In Formula (23), $\sigma(\cdot)$ indicates sigmoid function. In order to avoid losing the good properties of the original features, the features weighted by the attention mask are designed as Formula (24):

$$F_j = (1 + m) \otimes G_j \ , \tag{24}$$

where $\otimes$ indicates that the corresponding elements are multiplied. This similar residual connection method not only effectively retains the backbone characteristics, but also superimposes the significant characteristics. It is worth noting that with the deepening of dense connection convolution module, spatial attention units at different levels can capture different perceptual attention from local to global, so as to more accurately locate the expression region of interest.

3 Output layer

In this paper, multiple densely connected convolution modules and spatial attention units work together to extract emotion rich spatial feature vector f. Due to the uneven differences of students' facial structure and expression actions in the same expression samples, the similar expression features are sparsely distributed in the feature space, which affects the network's ability to distinguish. Therefore, multiple loss terms are used to control the feature distance, and the loss function contains two contents, that is, the cross entropy (CE) is used to minimize the classification error and the in class distance penalty term is used to reduce the change in the feature class.

The first part is the classification loss used to minimize the classification error. The classical cross entropy (CE) loss is used. The calculation expression is as Formula (25):

$$CELoss = -\sum_{i \in N} y_i \log \hat{y}_i \ , \tag{25}$$

where N represents the number of samples, y_i represents the probability distribution of the i-th sample's real label, and $\hat{y}_i$ represents the probability distribution of the i-th sample's predicted label, which is calculated as Formula (26):

$$\hat{y}_i = softmax(W^o f_i + b^o) \ , \tag{26}$$

where $Softmax(\cdot)$ represents the normalized exponential function, W^o and b^o represent the parameter matrix and offset of the full connection layer, and f_i is the spatial eigenvector of the i-th sample. The classification loss monitoring model learns the distinguishing features of different classes of samples. The second part is the intra class distance penalty term used to reduce the variation within the feature class. The calculation expression is as Formula (27):

$$PELoss = \frac{1}{2} \sum_{i \in N} max\{ \| f_i - z_{c_i} \|_2^2 - d_{c_i}^{\ 2}, 0 \} \ , \tag{27}$$

where, c_i represents the category of the i-th sample, z_{c_i} is the mean value of eigenvectors of all c_i samples, $\|f_i - z_{c_i}\|_2$ calculates the L2 norm of f_i and z_{c_i}. That is, the Euclidean distance between the two vectors, d_{c_i} represents an in class distance constraint, and the calculation expression is as Formula (28):

$$d_{c_i} = \frac{1}{2}\min \| z_{c_i} - z_{c,c\neq c_i} \|_2. \quad (28)$$

In the process of reverse training, the partial derivative of $PELoss$ to f_i can be calculated as Formula (29):

$$\frac{\partial PELoss}{\partial f_i} = \begin{cases} f_i - z_{c_i}, if \| f_i - z_{c_i} \|_2 > d_{c_i} \\ 0, if \| f_i - z_{c_i} \|_2 \leq d_{c_i} \end{cases}. \quad (29)$$

The intra class distance penalty term achieves feature compactness through dynamic class center constraints: in each iteration, the running mean of the feature vectors of each expression category in the current batch is first calculated as the class center, and then the distance between each sample feature and the corresponding class center is calculated. When the distance is less than the preset threshold, the gradient is back-propagated to the feature extraction layer to drive the feature vectors to converge towards the class center, while samples exceeding the threshold do not participate in the constraint; By maintaining inter class differences, the intra class distribution variance of similar features is reduced, significantly improving the discriminative power of the feature space.

The penalty term can narrow the features with large intra class differences and limit the inconsistency of the same kind of samples within a certain local range. When the distance between the sample feature *i* and the corresponding class center is greater than the limit of the in class distance constraint, the feature i is not subject to the in class constraint. When the distance between the sample feature i and the corresponding class center is less than the limit of the intra class distance constraint, the feature i is punished so that the feature can approach the class center. The loss function is $Loss = CELoss + \lambda PELoss$, where λ is the hyper parameter, regulates the balance between the two terms. The network model boasts several benefits, including the promotion of feature reuse, the facilitation of feature propagation (or transfer), a reduction in computational overhead, and the mitigation of gradient vanishing issues.

5. Experimental Results and Analysis

5.1. Data Set

The experiment was verified on CK+, FER2013 and the self built student emotion database to evaluate the performance of the static expression recognition model. The database includes disgust, happiness, surprise and other expressions. Aiming at the problem that the sample image size is inconsistent but the model input size is fixed, the image size is adjusted to 224*224 by bilinear interpolation, and processing methods such as rotation, horizontal flip and illumination are used to amplify the data in the training process, so as to avoid the over fitting problem caused by too few samples in the data set.

The process of building a self built student emotional database is as follows: first, convene students from different grades as subjects; Design a multi scenario data collection plan for classroom interaction, experimental testing, etc., using high-definition cameras to record natural facial expression changes; Pixel level annotation of expressions such as anger, pleasure, and confusion; By cleaning the data to remove fuzzy samples, a dataset is constructed that covers different lighting, posture, and occlusion scenarios. Finally, cross validation is used to ensure annotation consistency, providing high ecological validity emotional data support for model training.

5.2. Development Environment

The system development language: Python 3.8. Development tools: Pycharm2020.1 and QT Designer, a

Table 1
Hyperparameter values.

Parameter	Value
learning rate	10^{-4}
batch	32
optimizer	Mini-Batch SGD
Dropout rate	0.1
weight decay	e^{-3}
Conv Kernel Size	7×7
epoch	150
padding	3
activation function	Sigmoid

visual GUI development tool. Framework and modules: computer vision and machine learning software OpenCV, deep learning framework Tensorflow, Tensorboard, numerical calculation module Numpy, and numerical analysis module Pandas. GPU for training: GeForce MX150. The initial learning rate is set to 10^{-4} and dynamically adjusted using cosine annealing strategy, decaying to 10^{-5} every 30 epochs to balance convergence speed and stability. Divide the training set, validation set, and testing set in a ratio of 7:2:1. Table 1 lists the hyperparameter values corresponding to model training.

5.3. Ablation Experiments and Model Comparison

To verify the impact of spatial attention mechanism on the model, we denote the lightweight densely connected convolutional neural network, the lightweight densely connected convolutional neural network based on spatial attention, and the lightweight densely connected convolutional neural network based on channel attention as LDCNN, SLDCNN, and CLDCNN, respectively.

Table 2
Model ablation experiment.

	CK+ accuracy	CK+ F1 score	FER2013 accuracy	FER2013 F1 score
LDCNN	95.2%	0.94	70.1%	0.68
CLDCNN	95.8%	0.95	71.5%	0.70
SLDCNN	96.3%	0.96	72.8%	0.71

Table 2 shows the impact of spatial attention mechanism and channel attention mechanism on the model on the CK+ and FER2013 datasets. After adding the spatial attention mechanism or channel attention mechanism, the accuracy on both datasets has improved to some extent. Specifically, the accuracy on CK+ has increased by 1.1% and 0.6% respectively with the spatial attention mechanism and channel attention mechanism, while on FER2013, the accuracy has increased by 2.7% and 1.4% respectively. This demonstrates that attention extraction of features highly related to emotion is beneficial for expression recognition.

The spatial attention mechanism focuses on the spatial locations of the feature map, such as the eye and mouth regions, and the generated key regions effectively cover the action units corresponding to expressions. For instance, when recognizing disgust, the model locates the key regions as the eyebrows, nose, and chin. When recognizing happiness, the model focuses on the cheeks and mouth. The channel attention mechanism focuses on the channel dimension of the feature map, such as color and texture channels. For micro-expressions (such as slight frowns and twitching of the mouth corners), spatial attention captures local changes through high-resolution feature maps, while channel attention may lose details due to global average pooling.

Spatial attention mechanisms exhibit significant advantages in facial expression recognition, thanks to their precise localization of key regions, strong anti-interference capabilities, high interpretability, and excellent computational efficiency. They demonstrate notable strengths, particularly when dealing with subtle expression changes and complex scenarios. In contrast, channel attention is more suitable for tasks dominated by global features. However, in facial expression recognition, which is sensitive to local features, spatial attention proves to be more competitive.

Table 3
Comparison of different network depths.

Network depths	CK+ accuracy	CK+ F1 score	FER2013 accuracy	FER2013 F1 score
12	95.2	0.94	68.4	0.62
20	96.3	0.95	70.1	0.65
32	96.7	0.96	71.8	0.67
44	96.5	0.95	71.2	0.66

Table 3 shows the comparison of different network depths. The CK+ dataset is a small-scale, controlled environment where performance peaks when the network depth is increased to 32 layers, achieving an accuracy of 96.7%. However, at 44 layers, there may be a slight decrease due to vanishing gradients or overfitting. The FER2013 dataset is a large-scale, natural scene environment where performance monotonically improves with depth. At 32 layers, the F1 score is 0.67, but it decreases at 44 layers, indicating that deeper networks are needed in natural scenes but redundant parameters must be avoided.

In table 4, we compare different models on two datasets using 10-fold cross-validation. Although the

Table 4
Model Comparison.

Verification method	CK+ accuracy	FER2013 accuracy
DCMA-CNNs	93.46%	69.4
MSCNN	95.54%	70.9
DAM-CNN	95.88%	71.2
ResNet-18 + CBAM	97.2	73.0
EfficientNet-B2	96.5	71.5
SLDCNN	96.7	71.8

methods for dividing the training and testing sets are slightly different, the accuracy still intuitively reflects the recognition capabilities of each model. DCMA-CNNs feature an integrated network of global and local feature branches. MSCNN introduces an expression verification signal into the loss function to supervise the distance between features. DAM-CNN is a single-layer attention CNN that eliminates information unrelated to expressions through an encoder-decoder framework. ResNet-18+CBAM combines a residual structure with an attention mechanism to enhance key feature extraction through channel and spatial attention, balancing lightweight and high accuracy. EfficientNet-B2 achieves a balance between high performance and low parameters by simultaneously optimizing depth, width, and resolution through a compound scaling strategy.

The SLDCNN model proposed in this paper achieves feature reuse through a multi-layer densely connected structure, and combines a spatial attention mechanism to precisely locate key facial expression regions (such as the corners of the mouth and the eyebrows). It achieves accuracies of 96.7% and 71.8% on the CK+ and FER2013 datasets, respectively, with only 1.78M parameters, significantly lower than ResNet-18+CBAM (11.7M) and EfficientNet-B2 (9.2M). Compared to the complex dual-branch structure of DCMA-CNNs, its single-branch densely connected design is more concise and efficient. Unlike the design of MSCNN, which relies on additional supervisory signals, our model naturally enhances feature discrimination through an attention mechanism. Compared to the single-layer attention of DAM-CNN, our model exhibits a more comprehensive multi-layer feature fusion capability. The automatically scaled composite coefficients of EfficientNet-B2 overly focus on texture features (such as wrinkles) in facial expression recognition tasks, neglecting key dynamic regions (such as the eyes). Experiments show that this model outperforms most existing methods in terms of lightweight design and performance balance, and with its small parameter count, it is suitable for real-time facial expression recognition scenarios on mobile devices.

Figure 5 shows the loss function and the accuracy of test set in neural network training visually based on CK+ data set.

Figure 5
Visual display of loss function and accuracy of CK+ dataset.

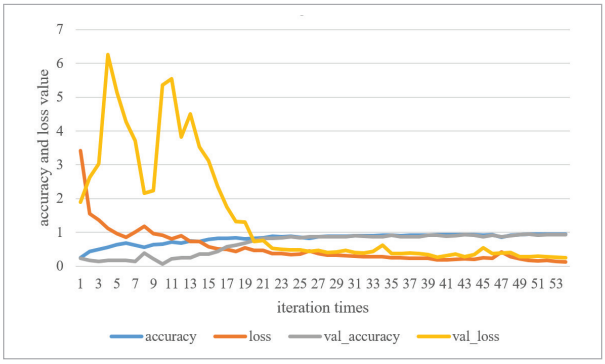
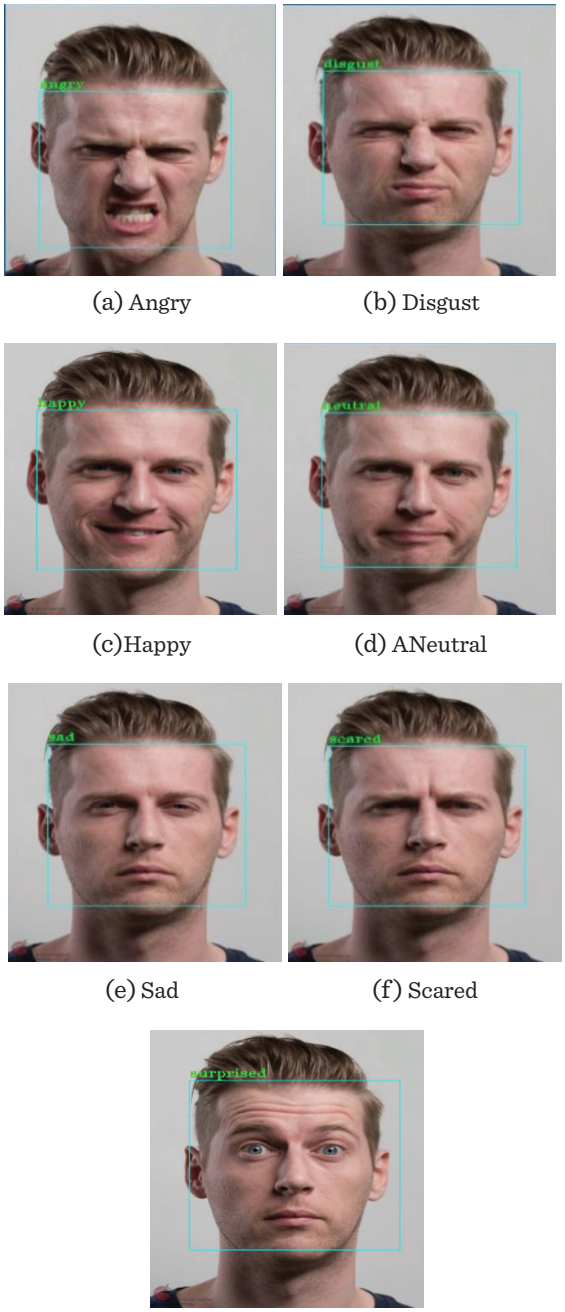


Table 5
The probability of facial expression classification predicted by the model in Figure 6.

Category	(a)	(b)	(c)	(d)	(e)	(f)	(g)
Angry	97.51%	25.39%	0.00%	1.01%	10.51%	34.02%	14.22
Disgust	0.79%	58.93%	0.00%	0.01%	0.02%	0.14%	0.00
Scared	1.07%	3.81%	0.00%	0.16%	7.46%	45.94%	11.65
Happy	0.00%	0.07%	99.40%	13.44%	0.00%	0.00%	0.01
Sad	0.63%	6.36%	0.00%	0.99%	54.34%	17.71%	13.76
Surprised	0.00%	0.03%	0.02%	0.03%	0.50%	0.21%	44.46
Neutral	0.00%	5.42%	0.57%	84.36%	27.16%	1.99%	15.90

Figure 6
Recognition results of different expressions.



(a) Angry (b) Disgust

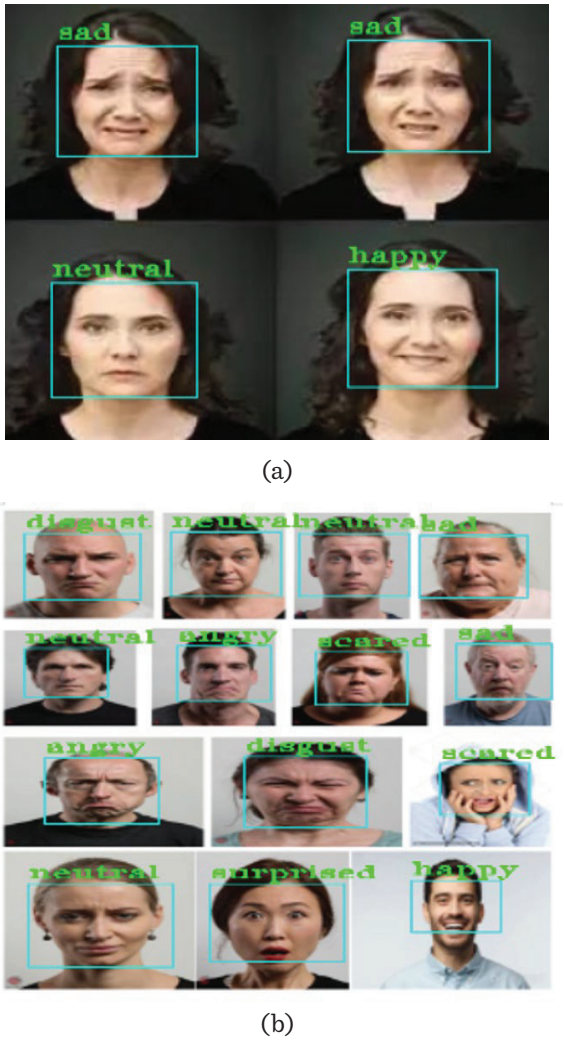
(c)Happy (d) ANeutral

(e) Sad (f) Scared

(g)Surprised

Figures 6-7 are static expression recognition test results. They respectively show the test results of single face image recognition and the expression recognition

Figure 7
Multiple facial expression recognition results.



(a)

(b)

Table 6
Percentage of people with multiple facial expression recognition in Figure 7.

Category	(a)	(b)
Angry	0.00%	14.29%
Disgust	0.00%	14.29%
Scared	0.00%	14.29%
Happy	25.00%	7.14%
Sad	50.00%	14.29%
Surprised	0.00%	14.29%
Neutral	25.00%	28.57%

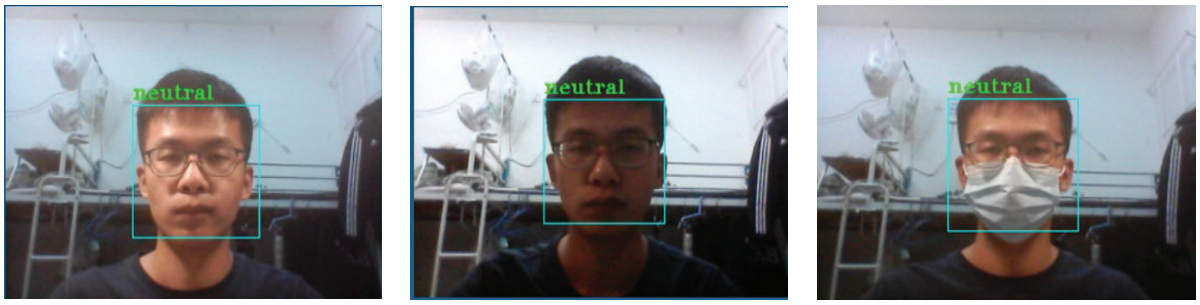
Table 7
The probability of facial expression classification predicted by the model in Figure 8.

Category	(a)	(b)	(c)
Angry	7.39%	15.24%	22.95%
Disgust	0.03%	0.02%	0.00%
Scared	4.69%	10.49%	7.98%
Happy	0.07%	0.25%	1.53%
Sad	9.68%	11.89%	26.91%
Surprised	0.20%	18.17%	0.23%
Neutral	77.93%	43.92%	40.40%

Table 8
Percentage of people identified by the model in Figure 9.

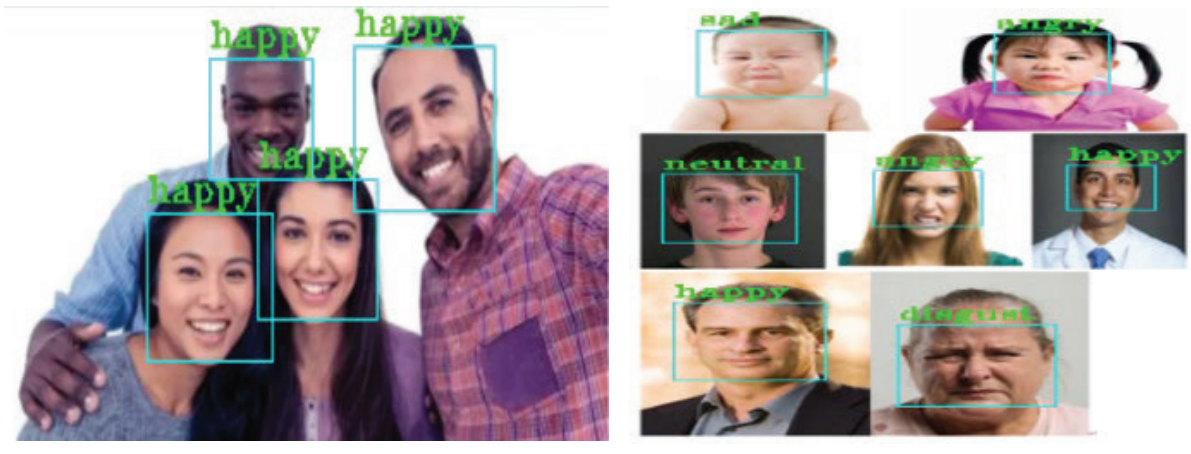
Category	(a)	(b)
Angry	0.00%	28.57%
Disgust	0.00%	14.29%
Scared	0.00%	0.00%
Happy	100%	28.57%
Sad	0.00%	14.29%
Surprised	0.00%	0.00%
Neutral	0.00%	14.29%

Figure 8
Detection results of normal indoor environment, low light intensity and face occlusion.



(a) (b) (c)

Figure 9
Recognition results of different skin colors and ages.



(a) (b)

test results of multiple face images. Table 5 is the probability of facial expression classification predicted by the model in Figure 6. Table 6 is the percentage of people with multiple facial expression recognition in Figure 7.

Figures 8-9 are dynamic expression recognition test results. Figure 8 shows the recognition and judgment results under the three conditions of normal indoor environment, low light intensity and face occlusion. Table 7 shows the probability of facial expression classification predicted by the model in Figure 8. It can be seen that there are certain differences in the accuracy of the same expression under different illumination intensity. When the illumination intensity is good, the accuracy of expression recognition is also high. When the illumination level is low, while facial expressions can still be detected, the recognition accuracy tends to decrease. Whether to wear a mask or not also has a small impact on the results. Facial coverings such as glasses and masks can obstruct the process of extracting facial feature values, resulting in the decline of recognition accuracy.

Figure 9 shows the recognition test results of different ethnic backgrounds and different age stages. Since most of the training sets use pictures of white and black people, the accuracy of identifying these two kinds of people in the test will be higher. But overall, these differences are not significant. The test results of face pictures at different ages are shown in Figure 9(b). Table 8 shows the percentage of people with different skin colors and ages.

6. Conclusion

This paper proposes a lightweight densely connected convolutional neural network based on multi-level spatial attention, considering the insufficient attention of convolutional neural networks to emotion-related local regions and the inability of classical loss functions to handle intra-class differences in expressions. Firstly, a densely connected convolutional module is employed to extract image features, obtaining rich information flows at multiple scales while controlling the number of model parameters. Then, based on the characteristics of the output features from the dense block, a spatial attention unit with minimal parameter count is constructed to focus on emotion saliency information. On this basis, an intra-class distance penalty term is introduced and combined with the classification loss to supervise parameter training, aiming to reduce the impact caused by differences in sample identity characteristics. The proposed model is validated on a dataset, and the experimental results effectively demonstrate the robustness and efficiency of the model.

The model presented in this paper demonstrates the advantages of efficient feature reuse and attention mechanisms in facial expression recognition. However, it still has two limitations: firstly, it is sensitive to extreme lighting conditions. Since the model relies partially on texture features, strong light or shadows can easily interfere with local keypoint detection. Secondly, due to the dense connection structure, it may overfit common facial expression patterns, resulting in fluctuations in performance on small sample categories. Future work can be focused on the following directions: 1) Introducing a multi-view fusion strategy, combining infrared or depth images to enhance robustness to lighting conditions; 2) Constructing a spatiotemporal joint model, such as embedding a temporal attention module, to leverage the continuity of video sequences to improve micro-expression recognition capabilities; 3) For the small sample problem, exploring meta-learning frameworks or Generative Adversarial Networks (GAN) for data augmentation to further balance category distribution.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, author-ship, and/or publication of this article.

Data Sharing Agreement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Acknowledgement

This work was supported in part by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant: 23KJA520013) the Xuzhou Science and Technology Plan Project (Grant: KC22305), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant: 22KJA520012), the sixth "333 project" of Jiangsu Province.

References

1. Adithya, A., Anisa, H., Monika, J. S., Thomas, T. S., Kumar, A. Facial Expression Detection and Classification Using SVM, CNN and Decision Tree Algorithm. International Journal of Health Sciences, 2022. https://doi.org/10.53730/ijhs.v6nS4.9770
2. Aouani, H., Ayed, Y. B. Deep Facial Expression Detection Using Viola-Jones Algorithm, CNN-MLP and CNN-SVM. Social Network Analysis and Mining, 2024, 14(1). https://doi.org/10.1007/s13278-024-01231-y
3. Arafin, S., Ashrafi, A. F., Alam, M. G. R., Talukder, A. BFER-Net: Babies Facial Expression Recognition Model Using ResNet12 Enabled Few-Shot Embedding Adaptation and Convolutional Block Attention Modules. IEEE Access, 2025, 13, 37302-37317. https://doi.org/10.1109/ACCESS.2025.3545759
4. Azrien, E. A., Hartati, S., Frisky, A. Z. K. Regularized Xception for Facial Expression Recognition with Extra Training Data and Step Decay Learning Rate. IAES International Journal of Artificial Intelligence, 2024, 13(4), 4703-4710. https://doi.org/10.11591/ijai.v13.i4.pp4703-4710
5. Badhe, S., Chaudhari, S. Deep Learning Based Facial Emotion Recognition. ITM Web of Conferences, 2022. https://doi.org/10.1051/itmconf/20224403058
6. Beibei, L., Jiansheng, Z., Suwen, L., Linlin, D., Zhiyuan, Y., Liangde, M. Real-Time Facial Expression Recognition on Res-MobileNetV3. China Communications (English Version), 2025, 22(3), 54-64. https://doi.org/10.23919/JCC.ja.2023-0123
7. Berrahal, M., Azizi, M. Augmented Binary Multi-Labeled CNN for Practical Facial Attribute Classification. Institute of Advanced Engineering and Science, 2021(2). https://doi.org/10.11591/ijeecs.v23.i2.pp973-979
8. Chandrasekaran, S. T., Jayaraj, A., Karnam, V. E. G., Banerjee, I., Sanyal, A. Fully Integrated Analog Machine Learning Classifier Using Custom Activation Function for Low Resolution Image Classification. IEEE Transactions on Circuits and Systems I: Regular Papers, 2021, 68(3), 1023-1033. https://doi.org/10.1109/TCSI.2020.3047331
9. Ciobanu, A., Shibata, K., Ali, L., Rioja, K., Andersen, S. K., Bavelier, S. K., Bediou, B. Attentional Modulation as a Mechanism for Enhanced Facial Emotion Discrimination: The Case of Action Video Game Players. Cognitive, Affective & Behavioral Neuroscience, 2023. https://doi.org/10.3758/s13415-022-01055-3
10. Deng, L. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. IEEE Signal Processing Magazine, 2012, 29(6), 141-142. https://doi.org/10.1109/MSP.2012.2211477
11. Dewi, C., Gunawan, L. S., Hastoko, S. G., Christanto, H. J. Real-Time Facial Expression Recognition: Advances, Challenges, and Future Directions. Vietnam Journal of Computer Science (World Scientific), 2024, 11(2). https://doi.org/10.1142/S219688882330003X
12. Fan, X., Jiang, M., Yan, H. A Deep Learning Based Light-Weight Face Mask Detector with Residual Context Attention and Gaussian Heatmap to Fight Against COVID-19. IEEE Access, 2021, 99. https://doi.org/10.1109/ACCESS.2021.3095191
13. Gao, X., Wu, S., Hu, W. X. Lightweight Image Super-Resolution via Multi-Branch Aware CNN and Efficient Transformer. Neural Computing & Applications, 2024, 36(10), 5285-5303. https://doi.org/10.1007/s00521-023-09353-8
14. Goel, R., Vashisht, S., Susan, S. An Empathetic Conversational Agent with Attentional Mechanism. International Conference on Computer Communication and Informatics (ICCCI), 2021. https://doi.org/10.1109/ICCCI50826.2021.9402337
15. Goodfellow, I. J., Erhan, D., Carrier, P. L. Challenges in Representation Learning: A Report on Three Machine Learning Contests. International Conference on Neural Information Processing. Berlin: Springer, 2013, 117-124. https://doi.org/10.1007/978-3-642-42051-1_16
16. He, K., Zhang, X., Ren, S. Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society, 2016: 770-778. https://doi.org/10.1109/CVPR.2016.90
17. Khanum, S. N. A., Mummadi, U. K., Taranum, F., Ahmad, S. S., Khan, I., Shravani, D. Emotion Recognition Using Multi-Modal Features and CNN Classification. AIP Conference Proceedings, 2024, 3007(1), 13. https://doi.org/10.1063/5.0192751
18. Khanna, D., Jindal, N., Rana, P. S., Singh, H. Enhanced Spatio-Temporal 3D CNN for Facial Expression Classification in Videos. Multimedia Tools & Applications, 2024, 83(4). https://doi.org/10.1007/s11042-023-16066-6
19. Krizhevsky, A., Sutskever, I., Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. Proceedings of the 25th International Conference on Neural Information Processing Systems, 2012: 1097-1105.
20. Kusuma, G. P., Jonathan, J., Lim, A. P. Emotion Recognition on FER-2013 Face Images Using Fine-Tuned VGG-16. ASTES Journal, 2020, 6. https://doi.org/10.25046/aj050638
21. Liao, M., Li, Y., Gao, M. Feature Coding Method Based on Shared Weights Support Vector Data Description

for Face Recognition. Journal of Physics: Conference Series, 2021, 1955, Article ID 1012029. https://doi.org/10.1088/1742-6596/1955/1/012029

22. Li, J., Ding, F., Hayat, T. A Novel Nonlinear Optimization Method for Fitting a Noisy Gaussian Activation Function. International Journal of Adaptive Control and Signal Processing, 2022, 36(3), 690-707. https://doi.org/10.1002/acs.3367
23. Liu, Y., Ji, L. Ensemble Coding of Multiple Facial Expressions Is Not Affected by Attentional Load. BMC Psychology, 2024, 12(1), 14. https://doi.org/10.1186/s40359-024-01598-9
24. Lee, J. H., Song, K. S. Comparison and Analysis of CNN Models to Improve a Facial Emotion Classification Accuracy for Koreans and East Asians. International Journal on Advanced Science, Engineering & Information Technology, 2024, 14(3). https://doi.org/10.18517/ijaseit.14.3.18078
25. Mao, L., Yan, Y., Xue, J. H., Wang, H. Deep Multi-Task Multi-Label CNN for Effective Facial Attribute Classification. Affective Computing. IEEE Transactions on (T-AFFC), 2022, 13(2), 11. https://doi.org/10.1109/TAFFC.2020.2969189
26. Mazen, F. M. A., Nashat, A. A., Seoud, R. Real Time Face Expression Recognition Along with Balanced FER2013 Dataset Using CycleGAN. International Journal of Advanced Computer Science and Applications, 2021, 6. https://doi.org/10.14569/IJACSA.2021.0120617
27. Nisamudeen, H. P., Zhang, L. Deep Learning-Based Facial Expression Recognition. Machine Learning, Multi Agent and Cyber Physical Systems, 2023. https://doi.org/10.1142/9789811269264_0032
28. Pei, E., Guo, M., Berenguer, A. D., He, L., Chen, H. F. An Efficient Illumination-Invariant Dynamic Facial Expression Recognition for Driving Scenarios. IET Intelligent Transport Systems, 2025, 19(1). https://doi.org/10.1049/itr2.70009
29. Simonyan, K., Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. Computer Science, 2014, 52(3), 1-14.
30. Sabir, D., Hanif, M. A., Hassan, A., Rehman, S., Shafique, M. TiQSA: Workload Minimization in Convolutional Neural Networks Using Tile Quantization and Symmetry Approximation. IEEE Access, 2021, Article ID 3069906. https://doi.org/10.1109/ACCESS.2021.3069906
31. Tang, Y., Cui, H., Liu, S. Optimal Design of Deep Residual Network Based on Image Classification of Fashion-MNIST Dataset. Journal of Physics Conference Series, 2020, 1624, Article ID 052011. https://doi.org/10.1088/1742-6596/1624/5/052011
32. Tang, Y., Yi, J., Tan, F. Facial Micro-Expression Recognition Method Based on CNN and Transformer Mixed Model. International Journal of Biometrics, 2024, 16(5), 463-477. https://doi.org/10.1504/IJBM.2024.140771
33. Traoré, C., Pauwels, E. Sequential Convergence of AdaGrad Algorithm for Smooth Convex Optimization. Operations Research Letters, 2021, 49(4). https://doi.org/10.1016/j.orl.2021.04.011
34. Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., Zafeiriou, S. End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. IEEE Journal of Selected Topics in Signal Processing, 2017, 14(8), 1-9. https://doi.org/10.1109/JSTSP.2017.2764438
35. Upadhyay, D., Malhotra, S., Gupta, M., Yadav, D. Multi-Disease Classification and Prediction Using Dense Visual Attention Network and Residual Network for: A Computer Vision Approach. International Conference on Advanced Informatics for Computing Research. Springer, Cham, 2025. https://doi.org/10.1007/978-3-031-84062-3_16
36. Wattanakitrungroj, N., Wettayaprasit, W., Rujirapong, P. T. S. Face Mask Classification Using Convolutional Neural Networks with Facial Image Regions and Super Resolution. IAES International Journal of Artificial Intelligence, 2024, 13(2), 2423-2432. https://doi.org/10.11591/ijai.v13.i2.pp2423-2432
37. Xu, Q., Xu, Z., Chen, Z., Chen, Y., Wang, H., Tao, L. Learning Sparse Filters-Based Convolutional Networks Without Offline Training for Robust Visual Tracking. Applied Intelligence, 2025, 55(6): 1-23. https://doi.org/10.1007/s10489-025-06350-3
38. Zhao, X., Zhang, S. Facial Expression Recognition Based on Local Binary Patterns and Kernel Discriminant Isomap. Sensors, 2011, 11(10), 9573-9588. https://doi.org/10.3390/s111009573
39. Zhu, K., Du, Z., Li, W., Huang, D., Chen, L. Discriminative Attention-Based Convolutional Neural Network for 3D Facial Expression Recognition. 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). IEEE, 2019. https://doi.org/10.1109/FG.2019.8756524