Design and Research of Facial Expression Recognition System Based on Key Point Extraction

Yan Qu* and Yan Liu
Library, Yantai Vocational College, Yantai, 264670, China
[E-mail: quyan0605@outlook.com]
*Corresponding author: Yan Qu

Received October 31, 2023; revised March 10, 2024; revised May 10, 2024; accepted May 17, 2024; published January 31, 2025

Abstract

Currently, facial recognition is very common in the use of libraries, such as self-service borrowing systems and access gate systems. The face recognition system, which extracts facial key points, has become an integral part of people's daily lives and work. To enhance facial recognition accuracy, researchers have combined spatio-temporal graph convolutional network models with temporal feature information to extract facial expression features and recognize dynamic expressions through facial expression sequences. Additionally, they have introduced an adaptive attention mechanism and automatic adjustment of attention distribution for peak frame images, resulting in the acquisition of more comprehensive image information. The results indicated that the accuracy of the model increases by an average of 1.595% with the introduction of the adaptive module, resulting in a final recognition accuracy of 97.26%. Compared to the independent spatio-temporal graph convolutional network model, the accuracy was increased by an average of 7.65%. In conclusion, the proposed adaptive spatio-temporal graph convolutional network model based on peak frame image optimization has better facial expression recognition performance. This improved technology helps to improve the management efficiency of the library.

Keywords: Feature extraction, Image matching, Spatio-temporal map convolutional networks, Adaptive mechanisms, Peak frame, Data fusion

1. Introduction

Visual images are the most widespread means of disseminating information, so recognition technology for images is an important development trend in the digital field [1]. Facial recognition technology has convenience and scalability in intelligent libraries. Users only need to register once upon entering the library, input their facial information into the system, and then use facial recognition technology to automatically recognize and achieve fast passage, saving the trouble of carrying access cards or remembering passwords. At the same time, it can be integrated with other systems, such as borrowing systems and systems for finding lost books, providing more convenient functions and enhancing security [2]. Expression recognition of human faces can be divided into dynamic and static parts. Static expression recognition techniques are relatively more mature and can be divided into traditional methods and depth extraction methods. Among them, the traditional feature extraction methods are divided into texture feature extraction and geometric feature extraction according to the form of data [3]. Although the development of recognition techniques for static expressions is more far-reaching, in practice, expressions are not stable and unchanging, but more often a dynamic process of change. This means that features contain a large amount of temporal feature information, and general research has used expression sequences to achieve recognition of dynamic expressions. Commonly, feature fusion, sequence feature extraction and deep learning networks are used. One of these is deep network-based recognition techniques. Making adaptive extraction of spatio-temporal data possible is a major breakthrough in development [4]. Deep learning mimics the structure and function of the human brain by constructing and training artificial neural network models to learn and perform complex tasks from data. It is widely used in various fields such as computer vision and facial recognition, and has achieved significant results in tasks such as image classification, facial recognition, and machine translation. Since traditional face recognition methods mainly focus on static expressions, there is still room for improvement in recognizing dynamic expressions [5].

Dynamic expression recognition requires capturing facial movements that change over time. Spatio-temporal graph convolutional networks can capture features in both temporal and spatial dimensions. Compared to traditional methods, this model can better handle and learn the complexity of dynamic expressions. Therefore, this research combines spatio-temporal graph convolutional network models with temporal feature information to extract expression features, recognize dynamic expressions through expression sequences, and process dynamic expressions. Additionally, an adaptive attention mechanism and peak frame images are introduced to automatically adjust the distribution of attention, thereby obtaining more comprehensive image information. The research aims to improve the accuracy and robustness of facial recognition systems by addressing the issues of insufficient model generalization ability and insufficient preservation of detailed features. The technology's innovation lies in combining facial expression feature extraction and temporal feature information using a spatio-temporal graph convolutional network model in the recognition process. Additionally, an adaptive attention mechanism is introduced to fuse peak frame images with spatio-temporal maps while automatically adjusting attention distribution. This study aims to enhance the performance of facial recognition systems by focusing on dynamic expressions, which are essential for completing complex facial recognition tasks.

The article is structured into four main parts. The first part introduces the applications of image matching algorithms at home and abroad. The second part introduces the differences between the spatio-temporal map convolutional network model and the traditional

convolutional network, and introduces an adaptive attention mechanism and peak frame images to improve the model in the light of its still existing shortcomings such as poor generalization and insufficient retention of detailed features. The third section verifies the effectiveness of the improved spatio-temporal map convolutional network model using simulation experiments. And in the fourth part, a detailed summary of the experimental results data is presented, concluding that the adaptive mechanism network model fusing the peak frame image and the spatio-temporal map can achieve better recognition results compared with other recognition techniques.

2. Related Works

Face recognition is a major area of study in sentiment analysis that has drawn a lot of interest from academics both domestically and abroad. Both conventional and deep learning feature extraction approaches are generally included in the technique. Traditional approaches have drawbacks, though, and in modern studies pixel-value approaches are frequently integrated with other approaches. Sun Z et al. argued that classical feature extraction methods do not take into account the regional relevance of pixel values, and to further enable autonomous learning of this information, the study introduces an adaptive learning method. This approach required the formation of an intra-class low-rank dictionary from the original space, followed by the generation of an active feature dictionary before finally embedding a principal component module for filtering important information and dimensionality reduction. The method was used in simulation experiments on a known dataset and the experimental results show that the method has reliable performance for image recognition [6]. Iqbal MTB et al. argued that existing expression recognition techniques have not been able to achieve feature descriptors with more uniqueness as well as robustness. Based on this, the study proposed a novel feature descriptor for local shape patterns, which incorporates the gradient information of each neighborhood into the statistical range and describes its local shape using orientation feature data for the purpose of noise reduction and image smoothing. The selection strategy theory was also introduced to calculate the coding weights and the feasibility of the method was verified through experiments [7].

It can be noted that with further research, feature extraction methods based on deep learning networks are gradually becoming mainstream. liu C et al. argued that the traditional pixel-value algorithm, does not take into account the location features of key points of the face. The study therefore proposed a recognition method that combines convolutional neural networks with long and short-term memory networks, while introducing an adaptive attention mechanism to improve the generalisability of the model. Finally the study conducted simulation experiments on three datasets, FER2013, CK+ and JAFFE. And the experimental results revealed that the recognition performance of the method is significantly better than the other performance [8]. Ramachandran B et al. selected an integrated deep learning network as the expression recognition technique, which was divided into three main layers, namely the CDS feature vector layer, the expression probability vector layer and the meta-classifier layer. The CDS feature vector contained three main features: angle, length and slope. The final study used Bosphorus and the CASIA 3D database to simulate the model to verify the feasibility of the method [9]. Swaroop KV et al. proposed a convolutional neural network-based expression recognition technique using a sample of data collected by the network. It was compared with the classical pixel-value recognition algorithm and the experimental results showed that the recognition accuracy of the pixel-value algorithm and the convolutional neural network were 82% and 93% respectively.

Therefore, the convolutional neural network algorithm proposed in the study has better recognition results [10]. S.A.M. Al-Sumaidaee et al. applied spatio-temporal modeling to dynamic expression recognition, transformed spatial sequence images according to direction and angle, used robust gradient components to process recognition under low light intensity, and described dynamic expression changes by extending slender five-element pattern descriptors. The experimental results showed that in the MMI database, the recognition rate is as high as 79.23% [11]. H. Zaaraoui et al. applied minimum string to face recognition. This feature extraction method was not limited to the analysis of microstructure information, but also used mask surrounding to process pixels, which provides convenience for image coding. Finally, the effectiveness of this method was verified by experiments [12].

In summary, both convolutional neural network (CNN) and long short-term memory (LSTM) models can be applied to video analysis. Deep learning technology is gradually becoming a research hotspot and mainstream technology in the field of facial recognition. The test results of these methods on multiple datasets have demonstrated their superior recognition performance compared to traditional feature extraction techniques. However, CNN is primarily used for static images and has limited processing power for dynamic or sequential images, while LSTM has higher computational costs. Therefore, this research uses spatio-temporal graphical CNNs in the field of face recognition, while introducing adaptive attention mechanisms and the concept of peak frame images to improve the original model in terms of generalisability and extraction of detailed features.

A summary of existing research comparisons is shown in **Table 1**.

Table 1. Comparison of Existing Studies

Reference	Approach/ techniques	Performance and strength	Weakness
[6]	Adaptive learning	The reliability of image recognition	Complexity and increased computational costs
[7]	Number of directional features and selection strategy	Noise reduction and smooth image, feasibility	Specific parameter selection is required
[8]	CNN+LSTM+Adaptive Attention Mechanism	Improve the generalization of the model and achieve better recognition performance	The model is complex and there is overfitting
[9]	Integrated deep learning network	Feasibility	Contains a large number of parameters
[10]	CNN and network data collection	The recognition accuracy is 93%	Performance decreases under different lighting conditions

3. Design of A Face Recognition System Based on Expression Keypoint Extraction

Face recognition is widely used in various fields to help humans live intelligently and is one of the key trends in research today. The expression of the human face is a dynamic change process and therefore accompanied by a high temporal nature. The temporal features should be the main condition in the selection of the recognition network model, while the computational effort of the model should be minimized to achieve efficient work. The technical route of the study is shown in **Fig. 1**.

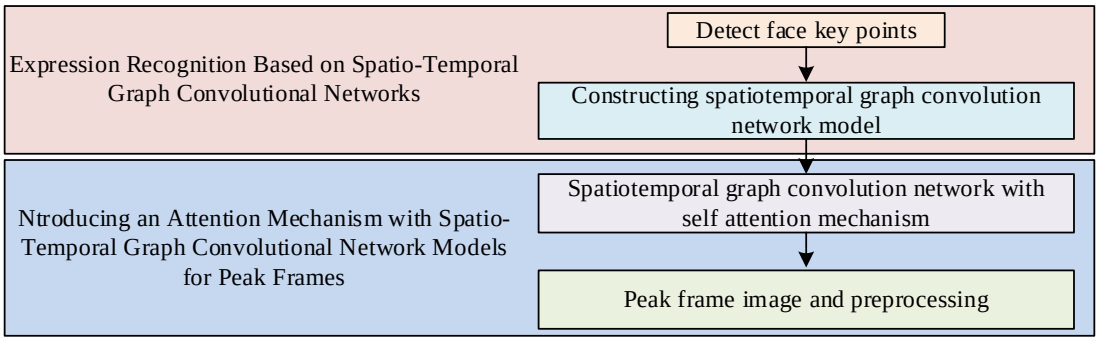


Fig. 1. Technical route

3.1. Expression Recognition Based on Spatio-Temporal Graph Convolutional Networks

By recognizing user facial expressions, intelligent systems can analyze their emotions and behaviors within the library, thereby optimizing the library's services and environment. Key point extraction of expressions refers to the analysis of the localization of key parts of a face image, such as the eyes, nose and lips, whose distribution is closely related to the organizational structure of the face and therefore contains a large amount of data information, which is a key component affecting the accuracy of the recognition model [13]. As the research continues to refine, the number of key point annotations has become roughly 500 times greater than the initial number, and the accuracy of the model has improved significantly [14]. The study chose the face key point annotation format from the Dlib open source library, an algorithm that gradually converges to the exact value based on residual regression, and includes a total of 68 points based on eye, nose, mouth and eyebrow as well as contour annotation. However, some scholars have analyzed that face contours actually have little effect on expressions [15]. Therefore, the study discarded 17 contour keypoints and only made the remaining 51 keypoints available for use in the model, the set of coordinates for all key points, as shown in equation (1):

$$LM = \left\{ LM_{t,i} \middle| 1 \le t \le S, 1 \le i \le N \right\} \tag{1}$$

In the above equation (1), $LM_{t,i} = \left(x_{t,i}, y_{t,i} \right)$ is the coordinate of the key point, t denotes the number of image frames, i denotes the key point number, and N is the total number of key points in the image taken here as 51. S is the frame length of the expression sequence; where the coordinates of the key point can be expressed as $(x_{t,i}, y_{t,i})$. The face key point annotation described above is shown in **Fig. 2**.

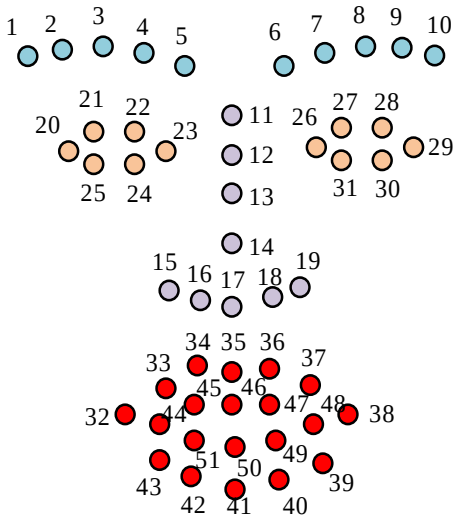
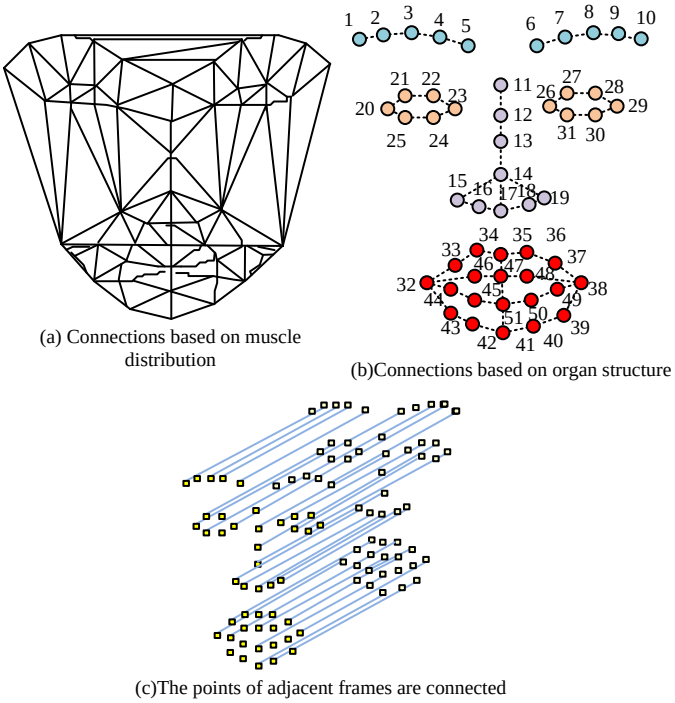


Fig. 2. Face key point annotation

The face is not static, but constantly changing, and the spatio-temporal graph convolutional network can extract such feature points containing temporal data, denoted by $G=(V,E)$ where G, V and E denote the spatio-temporal map, the node set and the edge set respectively. The node-set is the set of keypoints in all frames, while the edge-set contains the edge-set E_S between keypoints in a particular frame, and the edge-set E_T of keypoints between neighboring frame numbers [16]. The methods of connecting information between nodes can be divided into three forms: muscle distribution-based connection, organ structure-based connection, and full connection, as shown in **Fig. 3**.



(a) Connections based on muscle distribution

(b)Connections based on organ structure

(c)The points of adjacent frames are connected

Fig. 3. Connection modes of the same frame and adjacent frames

Fig. 3(a) and **3(b)** display the connection of various keypoints within the same frame. The former utilizes the Delaunay triangular dissection method to ensure that the connected edges are unique and do not interfere with each other. However, in dynamic expressions, the edge network changes slightly as the keypoints shift. Therefore, the study standardizes it into a static connection pattern. The complete connection is employed for the connection of keypoints at a distance between identical frames. There is a correlation between facial expressions and organs. For instance, when crying, both the lips and eyes turn downwards. In **Fig. 3(c)**, the keypoints between neighboring frames are connected, with the darker color representing the previous frame and the opposite color representing the next frame. Equation (2) shows the connection of all keypoints' edges.

$$\begin{cases} E_S = \left\{ v_{t,i} v_{t,j} \middle| (i,j) \subset H; i,j \in [1,N] \right\} \\ E_T = \left\{ v_{t,i} v_{(t+1),j} \middle| i,j \in [1,N] \right\} \end{cases} \tag{2}$$

In the above equation (2), H represents the set of key point numbers in different connection forms, $v_{t,i}v_{t,j}$ and $v_{(t+1),j}$ both represent two-dimensional variables. In the same frame edge set E_S, the two-dimensional variable is set to equal 1 when the two keypoints are connected with the value $(i,j) \subset H$, otherwise it is 0. In the neighboring frame edge set E_T, the two-dimensional variable is set to 1 when the two keypoints are equal, otherwise it is 0. As previously stated, facial expressions are a dynamic process, making the two-dimensional algorithm impractical for real-world applications. To address this issue, the study proposes the use of a spatio-temporal convolutional network to classify key points in dynamic images through the iterative fusion of information from each key point. **Fig. 4** illustrates the structure of the spatio-temporal graph convolution module.

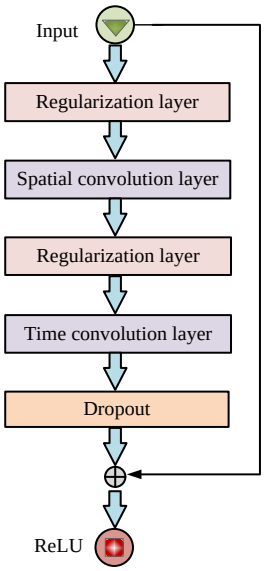


Fig. 4. Structure of spatio-temporal graph convolution module

Fig. 4 shows that the module is composed of four parts: a regularization layer, a time/space convolution layer, and a dropout layer. The regularization layer normalizes the feature points to retain the original feature data. The spatial and temporal convolution layers fuse node data between a frame and its neighbors, respectively. Additionally, the entire spatio-temporal

graph is compressed using a temporal convolution layer. The spatio-temporal graph convolutional network is utilized to perform convolution operations on each frame sequence and extract temporal features. The use of a dropout layer is an effective method to prevent overfitting problems in deep learning models, which improves model performance and robustness. Additionally, the generalization of the system is further enhanced through the processing of the dropout layer. When the dataset is not large enough, it may cause overfitting as well as feature convergence phenomenon. Jump connection can avoid this problem by performing non-linear transformation directly after feature summation [17]. The ReLU function introduces nonlinearity to enable the model to capture complex spatio-temporal data patterns and improve its expressive power. This ensures that the model can capture nonlinear relationships. CNNs typically use a two-dimensional input-output feature graph that is well-organized. However, the arrangement between nodes in the graph structure is irregular. To improve traditional convolutional networks, this study introduces the concept of sampling and weighting functions through convolution-like layers. Equation (3) demonstrates how the sampling function chooses a fixed range of neighborhoods centered at a point.

$$N(v_{t,i}) = p(v_{t,i}) = \left\{ v_{t,j} \middle| d(v_{t,i}, v_{t,j}) \le D \right\} \tag{3}$$

In the above equation (3), N is the set of neighborhood nodes and the range size of the neighborhood is $K \times K$. d refers to the minimum value of the distance between two nodes and D denotes the path, which generally takes the value of 1. Each key point in the sampled neighborhood range corresponds to a unique weight vector. In the graph structure, the set of neighborhoods N is divided into incompatible subsets of size M according to the distance between the points, and the nodes are labeled l according to the rules, and the dimensional value c of the weight vector is obtained from them, as shown in equation (4).

$$f_{out}(v_{t,i}) = \sum_{v_{t,j} \in N(v_{t,i})} \frac{1}{Z_{t,i}(v_{t,j})} f_{in}(v_{t,j}) \cdot w(l_{t,i}(v_{t,j})) \tag{4}$$

In the above equation (4), the mapping set is $l_{t,i} : N(v_{t,i}) \to \{0, ..., M-1\}$ and $Z_{t,i}(v_{t,j})$ is the total number of nodes in the neighborhood of a node that agree with the label of $v_{t,j}$, a value that balances the output weights of the nodes. Equation (5) expresses the neighborhood set and corresponding labels, allowing for smooth propagation of information to nodes in adjacent frames through the temporal map convolution layer. This layer also connects identical nodes between different frames.

$$\begin{cases} N(v_{t,i}) = \left\{ v_{q,i} \middle| |q-t| \le \left| \frac{m}{2} \right| \right\} \\ l_{t,i}(v_{q,i}) = l + q - t + \left| \frac{m}{2} \right| \end{cases} \tag{5}$$

In the above equation (5), m denotes the span of nodes of the same number; the size of the convolution kernel of the spatio-temporal graph is $[m \times 1]$ and the step size is s. The spatio-temporal graph convolutional network consists of seven spatio-temporal graph

convolutional modules, as well as three pooling, fully connected, and classification layers. The time span is set to 5 and the step size is set to 1, except for the 3rd and 6th modules, which are set to 2.

3.2. Introduction of an Attention Mechanism with Spatio-Temporal Graph Convolutional Network Models for Peak Frames

In the intelligent library user expression recognition model, spatio-temporal graph convolutional networks can reduce computational burden and achieve effective dimensionality reduction while solving the recognition problem of dynamic images, which is a significant improvement compared to traditional convolutional networks. However, the same topology can make the model dull and limit its generalization for independent data samples, which can seriously affect the model's ability to recognize expressions [18]. Therefore, the study introduces adaptive extrema to enhance the uniqueness of the feature data and help achieve better image classification results. The input quantity of this mechanism is the set of node features $h=\left\{\overrightarrow{h_1},\overrightarrow{h_2},...,\overrightarrow{h_N}\right\}$ and the output value is the new set of node features $h^{'}=\left\{\overrightarrow{h_1^{'}},\overrightarrow{h_2^{'}},...,\overrightarrow{h_N^{'}}\right\}$, as shown in equation (6).

$$\begin{cases} h=\left\{\overrightarrow{h_1},\overrightarrow{h_2},...,\overrightarrow{h_N}\right\} \\ h^{'}=\left\{\overrightarrow{h_1^{'}},\overrightarrow{h_2^{'}},...,\overrightarrow{h_N^{'}}\right\} \end{cases} \tag{6}$$

When the number of nodes is i, input set $\overrightarrow{h_i}\in R^F$, output set $\overrightarrow{h_i^{'}}\in R^{F^{'}}$, F and $F^{'}$ represent the number of input and output node features, respectively. This study combines input features and introduces a self-attention mechanism to account for the temporal dimension of the spatio-temporal graph. The flow of the self-attentive mechanism is shown in **Fig. 5**.

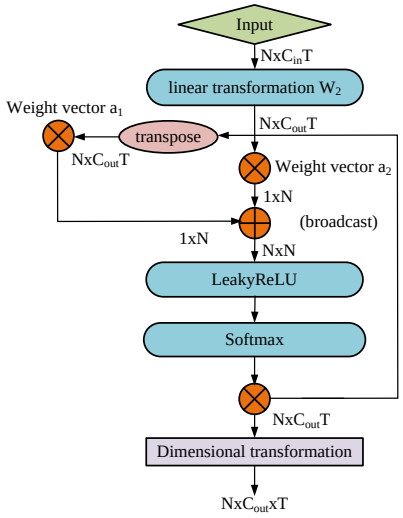


Fig. 5. Overall flow of self-attention mechanism

Fig. 5 shows that before introducing the attention coefficients, the original features should be transformed into higher order features. This enables the model to automatically adjust the attention distribution based on the content of each spatio-temporal graph, allowing it to focus on key frames with more expressive information. The study uses a linear transformation method to multiply the nodes with the weight matrix and then calculate the attention coefficients between the nodes as shown in equation (7).

$$e_{i,j} = a(W_2\overrightarrow{h_i}, W_2\overrightarrow{h_j}) \tag{7}$$

In equation (7) above, a denotes the attention mechanism $R^{F^{'}\times F^{'}} \to R$, $e_{i,j}$ denotes the attention coefficient between nodes i and j, and W_2 denotes the weight matrix. The Softmax function is utilized in the output layer to transform the output into a distribution that represents the probability of a class. This is especially appropriate for classification tasks. The Softmax function is then used to normalize the entire e values of node i and assuming the weight vector, the final attention coefficients are shown in equation (8).

$$a_{i,j} = \frac{\exp(\overrightarrow{a_1}W_2\overrightarrow{h_i})\cdot\exp(\overrightarrow{a_2}W_2\overrightarrow{h_j})}{\sum_{k=1}^{N}\exp(\overrightarrow{a_1}W_2\overrightarrow{h_i})\cdot\exp(\overrightarrow{a_2}W_2\overrightarrow{h_k})} = \frac{\exp(\overrightarrow{a_2}W_2\overrightarrow{h_i})}{\sum_{k=1}^{N}\exp(\overrightarrow{a_2}W_2\overrightarrow{h_k})} \tag{8}$$

Equation (8) shows that normalization results in node data loss. To prevent this, the study first applies the LeakyReLU function for nonlinear transformation. This allows negative inputs to have a non-zero gradient, reducing the risk of neuron death and avoiding the loss of node information. The study mainly uses ReLU and Leaky ReLU to increase network nonlinearity without changing feature representation scale. Finally, the individual coefficients are weighted and summed and widely applied to K attention mechanism as shown in equation (9).

$$\overrightarrow{h_i^{'}} = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{N}a_{i,j}^{k}W_2^{k}\overrightarrow{h_j}\right) \tag{9}$$

In the above equation (9), the value of a is usually set to 1, σ denotes the ELU non-linear conversion activation function, and its expression is shown in equation (10).

$$f(x) = \begin{cases} a(e^{x}-1), & x < 0 \\ x, & x \geq 0 \end{cases} \tag{10}$$

The ReLU function provides linear activation for positive input values and outputs 0 for negative input values. On the other hand, the ELU function has a mean output distribution of 0, which significantly reduces the training time of the model. And the broadcast mechanism addition method used in the study can convert the row vectors of matrices to the same dimension, which solves the problem that matrices of different sizes cannot be added together, as shown in equation (11).

$$\begin{bmatrix}1 & 2 & 3\end{bmatrix} + \begin{bmatrix}1\\2\\3\end{bmatrix} = \begin{bmatrix}1 & 2 & 3\\1 & 2 & 3\\1 & 2 & 3\end{bmatrix} + \begin{bmatrix}1 & 1 & 1\\2 & 2 & 2\\3 & 3 & 3\end{bmatrix} = \begin{bmatrix}2 & 3 & 4\\3 & 4 & 5\\4 & 5 & 6\end{bmatrix} \tag{11}$$

When $K=1$, the total number of nodes per image frame is N ,the temporal dimension is T , the feature dimensions of the input and output nodes are denoted as C_{in} and C_{out} , and the weight matrix $W_2 = R^{C_{in}T \times C_{out}T}$. The spatio-temporal convolutional network model now includes an adaptive mechanism that automatically learns and forms a unique topology based on differences in node features. This structure is then used to extract key features of different expressions, making them more easily recognizable to the model. **Fig. 6** illustrates the spatio-temporal graph convolutional network with the self-attentive mechanism introduced.

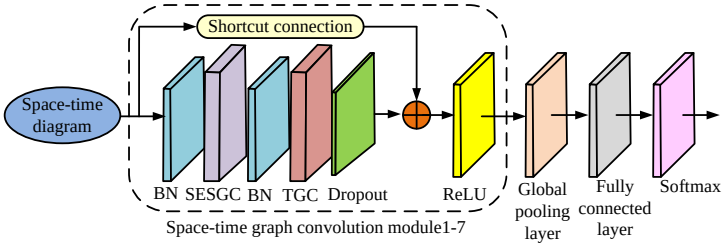


Fig. 6. Spatio-temporal graph convolutional network model with self-attention mechanism

Fig. 6 shows the improved model, which comprises six parts: the regularization layer, the self-attention enhanced spatial graph convolution layer (SESGC), the temporal graph convolution layer (TGC), dropout layer, ReLU, and shortcut connections. It can be seen that the temporal and spatial graph convolution layers are preceded by a regularization layer, then the final features formed by the dropout layer are added to the initial features, and finally a non-linear transformation is performed by the ReLU function. The module for spatial and temporal map convolution comprises the temporal and spatial map convolution layers, respectively, corresponding to their corresponding adaptive mechanisms. The purpose of the spatial graph convolution module is to identify the general features of the spatio-temporal graph. Then, in the self-attentive layer, it automatically learns how the points interact based on the feature values of each node, ultimately creating distinct features of various expressions [19-20]. Therefore, feature point updates in the spatio-temporal graph need to combine global features with the unique features of the image, and feature fusion is calculated as shown in equation (12).

$$f_{out}^{(l)} = f_g^{(l)} + a^{(l)} \bullet f_a^{(l)} \tag{12}$$

In the above equation (12), $f_{out}^{(l)}$ denotes the output features of the spatial convolution layer with the adaptive mechanism introduced, $f_g^{(l)}$ denotes the global features extracted by the spatial convolution layer. $f_a^{(l)}$ denotes the unique features of the image nodes extracted by the adaptive layer. The adaptive mechanism has significantly improved the model's sensitivity, allowing it to output unique features for different expressions. To further enhance the model's focus on local regions, sliding blocks are introduced to facilitate key local region learning. In contrast to the original subblock partition, this technique utilizes a sliding window approach to obtain more precise feature data from the high-level feature map. Each sliding window shares parameters to maximize efficiency. The overlapping regions between each slider enable multiple learning of the key block domain. The process of the sliding window is illustrated in **Fig. 7**.

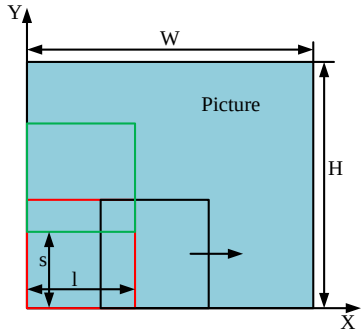


Fig. 7. Operation model of the slider attention mechanism

In **Fig. 7**, the size of the sliding block is l×l, the sliding distance is s, and both parameters belong to any positive integer. The sliding window first traverses all pixel values along the X-axis, then returns to the starting point, then glides s distance along the Y-axis, and traverses the X-axis again, and so on. Parameter constraints of sliding window are shown in equation (13).

$$\begin{cases} \left[1+\frac{W-l}{s}\right]=N, l\in N^{+}, s\in N^{+}, N\in N^{+} \\ 0<s<l \end{cases} \tag{13}$$

In the above equation (13), $\left[1+\frac{W-l}{s}\right]$ represents the maximum integer value not greater than the $1+\frac{W-l}{s}$ value. N indicates the total number of sliding Windows. The sliding block attention network is constructed using two bottleneck blocks, along with upper and lower branches. The bottleneck blocks preserve the size of the feature map while reducing its number. The upper branch of the network learns the eigenvector of the sliding block, while the lower branch learns the weight of the sliding block. The network runs in three steps. First, refine the feature map using bottleneck building blocks, which will serve as input data for the branches. Then, select the sliders in the upper branch and reduce dimensionality simultaneously. The weight of the sliding block is calculated in the lower branch. Then, the weight value is multiplied with the output feature vector to obtain the weighted feature of each window. However, the model's detection results are still affected by the environment's brightness and the algorithm's performance. Therefore, embedding improvements solely in the adaptive module is insufficient. The study introduces peak frame image theory to achieve better feature detection results. It emphasizes the importance of changes such as wrinkle texture of the face, which are as significant as temporal variation data in expression detection. The image uses a conventional CNN model to detect and extract texture features [21]. A suitable feature fusion method is employed to combine the peak frame image with the spatio-temporal map features. The peak frame refers to the state of the maximum number of frames from static to dynamic when the expression of a face occurs. The peak frames of dynamic sequences exhibit more pronounced expression features. The system under study can automatically identify these peak frames and assign higher attention weights to them, resulting in more effective capture of facial expression features. The detection process comprises four major steps: peak frame extraction, data pre-processing, feature extraction, and result classification. To avoid unnecessary computational burden and reduce computation speed, it is necessary to pre-process the data. This is because

experimental images often contain irrelevant elements such as background and hair, which can occupy a significant portion of the image. The study employs face detection from the Dlib open source tool library to isolate the facial region of the image. The study aimed to facilitate subsequent model input by unifying the size of intercepted face images to 224 x 224, as the focal length and size of each image varied. To extract peak frames, three CNN models were selected: the visual geometry group (VGG) network, Inception network, and ResNet network. The VGG network can have its number of layers configured according to requirements. Additionally, a 3x3 convolutional kernel is used, as the sum of small size convolutional kernels can achieve the same effect as large size convolutional kernels. This approach reduces the computational burden of the model and enhances its non-linear mapping effect. The Inception network's convolutional kernel is reduced to 1x1, and the model follows a classical construction method at the bottom, gradually evolving into modules such as Inception-v1 and Inception-v2 as the levels increase. This results in significantly improved performance. The ResNet network addresses the issue of gradient disappearance by summing multiple residual modules. The fusion of the three networks enables more efficient detection of peak frames. The study also introduces migration learning theory to prevent overfitting phenomenon [22]. The fusion of peak frame images with the output results of the spatio-temporal map is accomplished through the fusion of the feature layer and the decision layer, as illustrated in **Fig. 8**.

The study fuses the peak frame image with the last fully connected layer feature of the spatio-temporal map, as shown in equation (14).

$$(v_1, v_2, ..., v_N) \oplus (q_1, q_2, ..., q_M) = (v_1, v_2, ..., v_N, q_1, q_2, ..., q_M) \tag{14}$$

In the above equation (14), $(v_1, v_2, ..., v_N)$ and $q_1, q_2, ..., q_M$ denote the feature vector extracted from the spatio-temporal map and the feature vector extracted from the peak frame, respectively. N and M correspond to the dimensionality of the respective vectors. Decision layer fusion uses the classical weighted averaging method, where classifiers with different weights are summed to obtain the data classification results [23]. This is shown in equation (15).

$$l = w_0 \cdot p + w_1 \cdot q \tag{15}$$

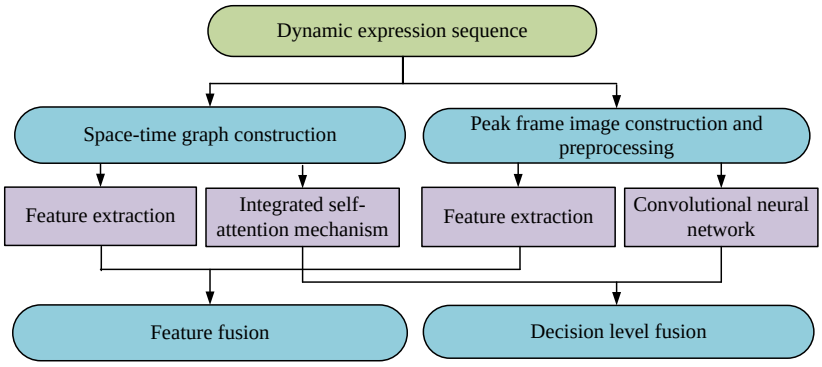


Fig. 8. Fusion of peak frame image and spatio-temporal graph output

In the above equation (16), p and q denote the output vectors of the temporal and peak frame maps respectively. w_0 and w_1 correspond to the weights of the two vectors respectively. All studies utilized a softmax classifier, with the output vectors representing probability values indicating the likelihood of a particular expression occurring [24-25]. The model used in this study is a spatio-temporal convolutional network with slider attention and peak frame images, as shown in **Fig. 9**.

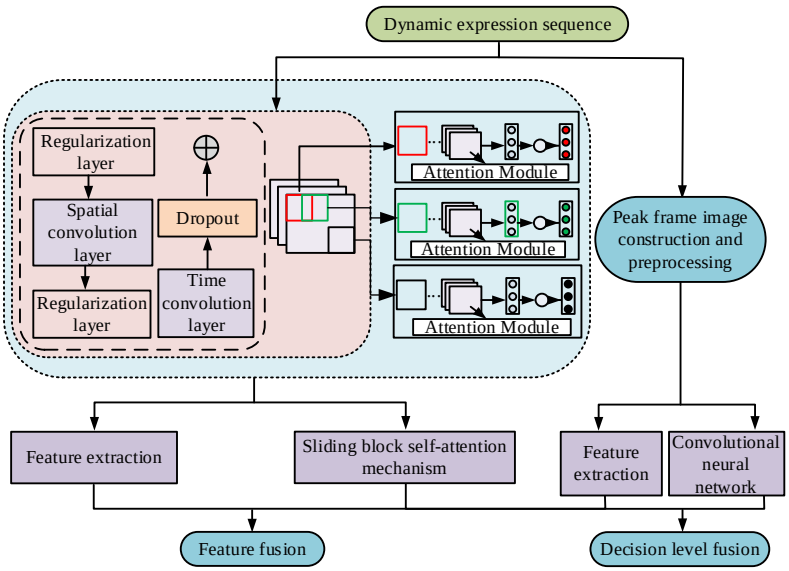


Fig. 9. Spatio-temporal convolutional network model based on sliding block attention and peak frame images

To begin with, the face expression's key points are extracted from the image using a specific algorithm. These points serve as essential indicators of facial expressions, including elements such as the eyes' positions and shapes, the brow region, and the lips. Afterward, a deep learning network called spatio-temporal graph convolutional network model, dedicated to processing temporal features, is applied. The sequence of key points is based on time series data input, and features are extracted by means of operations like convolution and pooling, enabling the identification of dynamic expressions. Secondly, to enhance the model's generalization skill and produce unique output features, the study introduces an adaptive sliding attention module, capable of automatically adjusting attention distribution for better focus and capture of key feature information. Finally, to enhance recognition accuracy, the study introduces the peak frame image fusion technology for fusing the peak frame image with the spatio-temporal map [26]. Choosing the image frame with the highest peak in the time series provides crucial temporal information. As a result, dynamic expressions can be well-represented, leading to improved accuracy in the model's recognition. In brief, the fundamental principles of the technology are extraction of expression key points, the use of the spatio-temporal graph convolutional network model, introduction of adaptive sliding attention module, and fusion of peak frame image and spatio-temporal graph. These principles are complementary and enable the technology to create an efficient face recognition system based on expression key points for intelligent libraries [27-28].

4. Simulation Experiment

The neural network construction in the study is done in the Pytorch framework using the Ubuntu 20.04 operating system. The study utilizes CK+ and Oulu CASIA datasets as experimental samples. The CK+ dataset contains 593 dynamic sequences from 123 volunteers, with 327 containing expression labels. The Oulu-CASIA dataset includes a total of 2880 dynamic sequences from 80 volunteers, with each image also featuring different imaging systems and versions of luminance. Thus, the selected dataset contains a total of 3473 dynamic sequences. It is important to note that samples with substandard luminance could not be used in the experiment, and therefore this type of sample should be removed directly as a preprocessing step. At the same time, to adjust the brightness level of the image, the dataset ensures similar brightness conditions for each image. This can be achieved through the histogram equalization method. The study divides the dataset into a test set and a train-test in a 1:9 ratio. In addition, the data preprocessing includes adjusting the sequence length to the same size value of 16, learning rates of 0.1 and 0.01 for the CK+ and Oulu-CASIA datasets respectively, batches of 100 and 256 respectively, training periods of 300 for both, and L2 regularization factors of 0.00001. The experimental setting is shown in **Table 2**. To assess the effectiveness and feasibility of the system, it is important to consider the hardware environment used in the experiment, including the processor type, graphics card specifications, memory size, and other related hardware devices. Additionally, it is necessary to analyze the performance test results of the subsequent system.

Table 2. Experimental environment parameter settings.

Hardware	Parameter
processor	Intel X eon(R)CPU E5- 2650 V4@2.20GHz
Video card	GeForce GTX 1080TI
memory	126GB
Hard disk	5.2TB SSD
Software	Parameter
CUDA	10.1.23
Cudnn	7.6.03
Pytorch	1.4.0
Python	3.6.12
OpenCV	4.5.1.48
Dlib	19.22.1

4.1. The Effect of Different Construction Methods on the Performance of Spatio-Temporal Map Convolutional Network Models

Considering the limitation of the sample data size, the study uses the ten-fold cross-validation method, dividing the dataset into 10 equal groups, and selecting one group of data in turn as the test set, while the remaining nine groups are used as the training set. The performance of the network models with different constructs in the two datasets is shown in **Fig. 10**.

Fig. 10 shows that the network models constructed from full connectivity achieve the highest recognition accuracy rates in both the CK+ and Oulu-CASIA datasets, reaching 93.91% and 78.73%, respectively. The network models based on muscle distribution and organ structure perform significantly worse than the fully-connected model. In the CK+ dataset, the recognition accuracy of the muscle distribution-based model is 1.57% lower than that of the fully connected model. In the Oulu-CASIA dataset, the model based on organ

structure performs 2.13% worse than the fully connected model. The fully connected construction network is 1.46% more accurate than the other two construction models in both data sets. It is evident that the two underperforming construction pathways have corresponding data dissemination deficiencies. For instance, the edge set based on muscle distribution only includes edges between each point and its neighbors. This results in a lack of smooth propagation of data information at longer distances when the number of model layers is insufficient. On the other hand, the organ-based network model cannot connect the node data between the subgraphs. Despite having a similar number of parameters, all three models performed worse when evaluated using the Oulu-CASIA dataset. This can be attributed to the limited number of expression categories in the Oulu-CASIA dataset, resulting in a significant difference in the overall number of parameters compared to the other dataset. A lower number of parameters leads to better network responsiveness. In summary, the study selects the fully connected form to obtain spatio-temporal information of dynamic images.

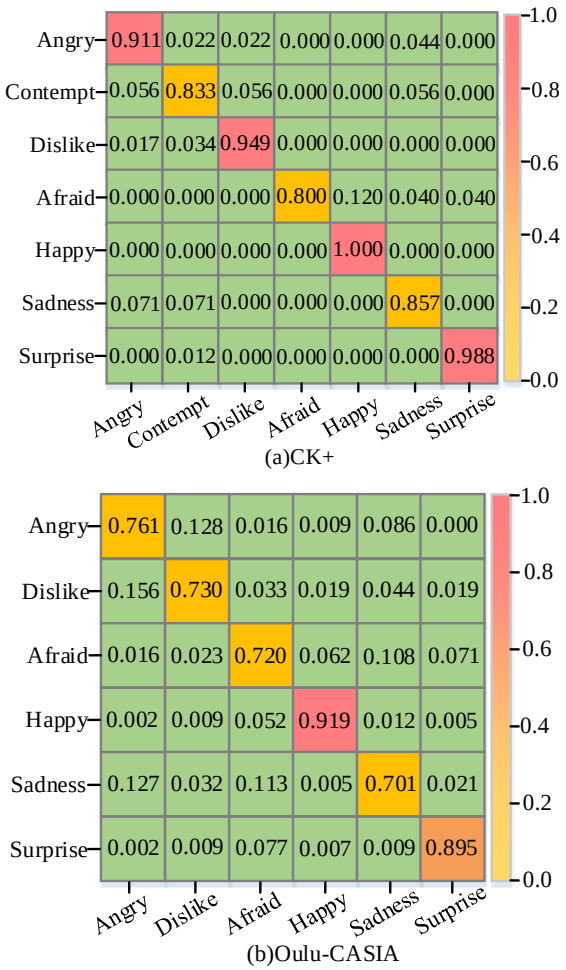


Fig. 10. Influence of construction mode on network model

4.2. Performance Implications of Convolutional Network Models with Embedded Adaptive Attention Modules for Spatio-Temporal Graphs

The study selects the cross-entropy function as the training loss function of the model, while the Adam optimization algorithm implements the model training process. Embedding the attention module with different number of heads into the spatio-temporal graph convolutional network model leads to the results shown in **Fig. 11**.

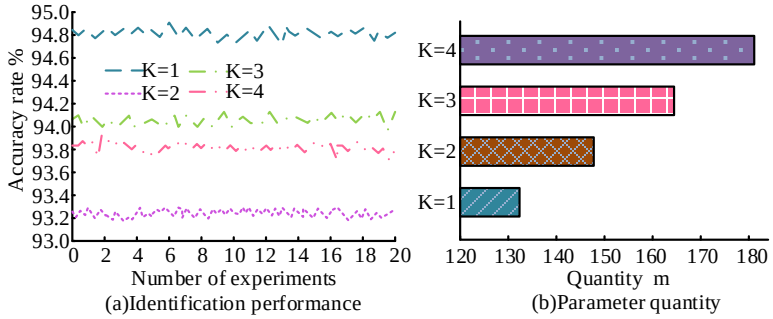


Fig. 11. Identification effect and number of model parameters on CK+ data set

Fig. 11 shows that the algorithm performs best with $K=1$, achieving a recognition accuracy of 94.79%, which is a 1.54% improvement compared to the least effective option. Additionally, $K=1$ has the smallest number of parameters, only 1,315,400. On the other hand, $K=4$ has 1,802,100 parameters, which is a relative increase of 486,700. The average accuracy of introducing various numbers of adaptive modules is 94.02%. Generally, increasing the number of self-attention mechanism heads can enhance the model's stability and learning ability. However, in CK+ datasets, increasing the attention module may lead to a decrease in recognition accuracy. This is because the dataset itself has a large sample content, and excessive addition of attention modules will increase complexity, ultimately leading to a decrease in recognition accuracy. **Fig. 11(b)** shows that the number of model parameters increases with the value. Therefore, the adaptive attention module of $K=1$ should be selected for the CK+ dataset. **Fig. 12** displays the performance in the Oulu-CASIA dataset.

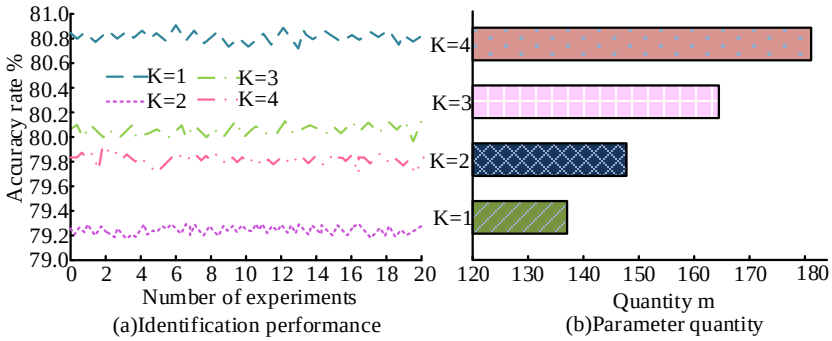


Fig. 12. Identification effect and number of model parameters on Oulu-CASIA data set

In **Fig. 12**, the model has the highest recognition accuracy when $K=4$, at 81.04%, which is a 1.85% improvement compared to the least effective. However, the number of parameters is also highest when $K=4$, at 1,812,000. The average accuracy for introducing different adaptive modules was 79.91%. In contrast to the CK+ dataset, the number of parameters in

this dataset increases as the number of attention modules increases, but the recognition accuracy does not decrease. Therefore, the parameter selection of the self-attention mechanism needs to be optimized based on the characteristics of the data set itself. The Oulu-CASIA dataset has a sufficient number of samples, so that the stability of the model's learning efficiency can be improved by adding adaptive modules as appropriate. Compared with the standalone spatio-temporal map convolution model, the introduction of the adaptive mechanism improved its real-perception accuracy by 0.93% and 2.31% in the CK+ dataset and the Oulu-CASIA dataset, respectively. Thus, the embedding of the adaptive attention module can well help the network model to achieve better recognition results. The weight values of the spatial map convolution layer and the self-attentive mechanism for feature fusion also play an important role. To test the effect of the weights on the performance of the algorithm, the study fixed the weights to 1 and continued the simulation experiments on the network model, as well as the weight calculation, to obtain the results shown in **Fig. 13**.

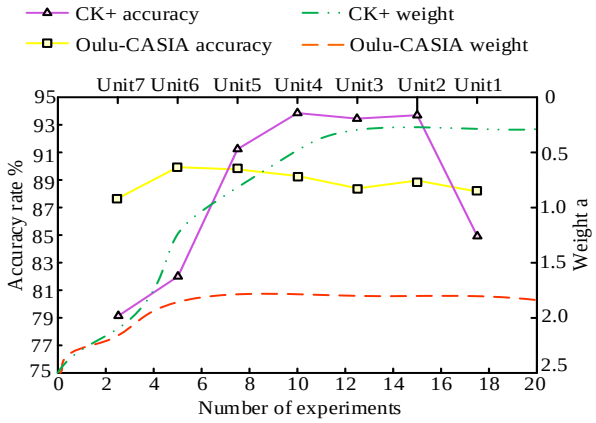


Fig. 13. Recognition accuracy of different data sets and weight of each module

To test the influence of the adaptive weights on the accuracy of the model, the weight value is set to 1. The dada line in **Fig. 13** shows the change in recognition accuracy of the two data sets with the number of experiments. When the fixed weight of 1 is adopted, the recognition accuracy of the model in the CK+ dataset and the Oulu-CASIA dataset reaches 93.26% and 80.19% respectively. It can be noted that the recognition effect of the model becomes worse under this weight value, decreasing by 1.53% and 0.85% respectively. Therefore, adaptive weights have an impact on model performance. In this paper, the tenfold crossover method is used to calculate the corresponding weight value of each module. The solid line in **Fig. 13** shows the weight changes of the two data sets. It can be seen that the weight of each module of the CK+ dataset is significantly different and unstable. In the second module the weight value has plummeted and in the fifth module it has always remained at an alternating low level. The difference in weight value between module 5 and module 6 is 1.2, and the average weight value of the high weight module is 1.63, which is larger than the initial training weight of 1. This indicates that in the CK+ dataset, the output features of the attention mechanism are more important than the output features of the spatial graph convolution layer. In the Oulu-CASIA data set, the change of its weight is relatively small, which shows a more stable change trend, and the maximum change of the total weight is not more than 0.5. Therefore, the output characteristics of the spatial graph convolution layer are more important in this dataset.

4.3. Performance Analysis of Convolutional Network Models Combining Peak Frame Images and Spatio-Temporal Maps

The study selected three models, VGG-16, Inception-v3 and ResNet-50, to implement the extraction of peak frame images, and experimentally compared each of these three types of models. The size of the input image is 224×224. To achieve optimal model recognition performance on the expression dataset, the size of the peak frame image after face detection must be converted to the corresponding size. The learning rates for the CK+ and Oulu CASIA datasets were set to 0.02 and 0.01, respectively, with batch sizes of 64 and training cycles of 150, and L2 regularization coefficients of 0.0001. The obtained results are shown in **Table 3**.

Table 3. Accuracy of different network models

CK+ data set			
Network model	VGG-16	Inception-v3	ResNet-50
Parameter quantity (m)	2270.21	2231.20	2403.39
Accuracy rate (%)	94.51	90.82	92.36
Oulu-CASIA data set			
Network model	VGG-16	Inception-v3	ResNet-50
Parameter quantity (m)	2270.13	2231.18	2403.39
Accuracy rate (%)	82.17	81.74	83.75

In **Table 3**, in CK+ dataset, the recognition performance of VGG-16 model is the best, and compared with Inception-v3 model and ResNet-50 model, the recognition performance is improved by 3.69% and 2.15% respectively. Moreover, the number of parameters of the three models is not much different, so in the CK+ dataset, VGG-16 model should be chosen as the basis of the sample peak frame image. In Oulu-CASIA dataset, the recognition performance of ResNet-50 model is the best, and the accuracy rate of RESNET-50 model is 1.58% and 2.01% higher than that of VGG-16 model and Inception-v3 model, respectively. The number of parameters is also reduced by 740,500 compared to the VGG-16 model. Therefore, in the Oulu-CASIA dataset, the ResNet-50 model should be selected as the basis for the sample peak frame images. At the same time, compared with the adaptive spatio-temporal graph convolutional network, the CNN has significantly more parameters, and its pre-trained convolutional kernel can help the network converge quickly under more parameters. And the convolutional network will redefine the classification layer, resulting in changes in the number of parameters. It is known that the fusion of peak frame images with spatio-temporal maps is divided into two aspects, the feature layer and the decision layer, and the study conducted simulation experiments on both separately. Among them, the experiments on the feature layer yielded the results as shown in **Fig. 14**.

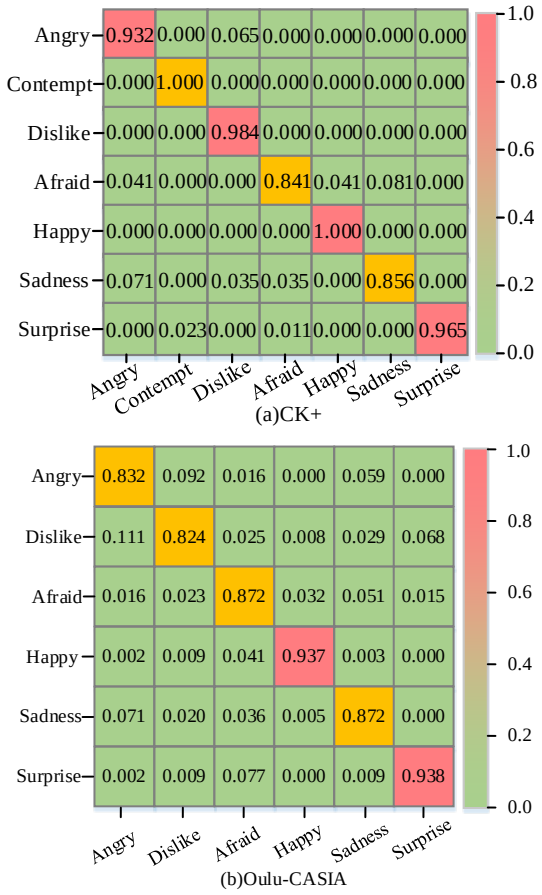


Fig. 14. Confusion matrix based on feature layer fusion model

In **Fig. 14**, the network model based on feature layer fusion has a recognition accuracy rate of 95.42% and 87.91% in the CK+ dataset and the Oulu-CASIA dataset, respectively. Compared with the spatio-temporal convolutional network model with adaptive mechanism, the recognition accuracy is increased by 1.53% and 6.87%, respectively. Compared with the peak frame graph and the spatio-temporal graph, the accuracy of the model has been significantly improved, making the optimization strategy of fusion of the two graphs feasible. Due to the different characteristics of the datasets, the recognition accuracy of the fusion models is also different. The fusion of the decision layer involves the selection of the w_0 and w_1 parameters and must satisfy the condition that they sum to a value of 1. The values of 0-1 were selected sequentially, resulting in the experimental results shown in **Table 4**.

Table 4 shows that in the CK+ dataset, the model achieves the best recognition accuracy of 97.26% with a w0 to w1 ratio of 0.1/0.9. This is a 2.46% and 2.76% improvement compared to the peak frame map and temporal map models alone, respectively. In the Oulu-CASIA dataset, the model achieves a recognition accuracy of 90.55% with a w0 to w1 ratio of 0.8/0.2. This is an improvement of 9.57% and 6.82% compared to the peak frame map and spatio-temporal map models alone, respectively. Therefore, fusing decision layers can enhance the model's recognition performance. The fusion model of the decision layer improves the correct rate by 1.83% and 2.64% respectively compared to the feature layer. This improvement is due to the fact that the graph itself has more repetitive data,

which the feature layer fusion does not effectively eliminate. The study compared the experimental data of different methods to further compare the impact of different methods on the model recognition accuracy, as shown in **Table 5**.

Table 4. Decision level parameter selection

CK+ data set			Oulu-CASIA data set		
Accuracy rate	w0	w1	Accuracy rate	w0	w1
94.81	1.0	0	80.97	1.0	0
95.42	0.9	0.1	89.53	0.9	0.1
93.89	0.8	0.2	90.55	0.8	0.2
96.97	0.7	0.3	89.82	0.7	0.3
96.01	0.6	0.4	89.45	0.6	0.4
96.01	0.5	0.5	87.82	0.5	0.5
96.34	0.4	0.6	89.04	0.4	0.6
96.65	0.3	0.7	89.68	0.3	0.7
96.32	0.2	0.8	88.41	0.2	0.8
97.26	0.1	0.9	88.52	0.1	0.9

Table 5. Data comparison between different methods

Method	CK+ data set	Oulu-CASIA data set
Muscle-based distribution	92.36%	77.73%
organostructure-based	92.65%	76.58%
Full connection	93.91%	78.73%
STGCN-SA	94.79%	81.04%
Feature layer fusion	95.40%	87.91%
Decision fusion	97.26%	90.53%

In **Table 5**, when only the spatio-temporal map model is used, the algorithm is significantly less effective in recognition than the other methods, and only has an advantage in the number of model parameters, which is less compared to the other methods. When a decision layer is used to fuse peak frame image features with spatio-temporal map features, the model has the strongest image feature recognition, reaching 97.26% and 90.53% in the CK+ dataset and the Oulu-CASIA dataset, respectively. This indicates that the recognition model based on peak frame images with spatio-temporal maps is better able to achieve correct matching. Therefore, the image recognition method proposed in the study is feasible and effective. The above experiments compare the influence of different modules in the model on the system identification performance, and verify the effectiveness of each module. To further verify the validity of the spatio-temporal graph CNN model based on adaptive mechanism and peak frame images proposed in this study, two datasets eNTERFACE05 and JAFFE are selected for training and verification. The recognition performance is compared with the CNN model proposed by B. Yang et al [29], and the graph CNN model proposed by R. Zhao et al [30]. Among them, CAER data set is the latest sentiment analysis data set, which belongs to a large database, with 1200 video data captured through TV video, respectively covering 44 different topics, and all videos are fully annotated data. This audio-visual data is more suitable for evaluating video feature extraction model. The JAFFE dataset has a relatively small sample size, consisting of only 213 data points for 10 individuals. In the previous and current experiments, the training, test, and verification sets included six basic emotions: happiness, sadness, disgust, fear, anger, and surprise. These emotions are distributed among the sets in a 70%, 20%, and 10% proportion, respectively.

Face pictures of different angles and image scales are selected as examples, and the process of face recognition by the model is shown in **Fig. 15**.

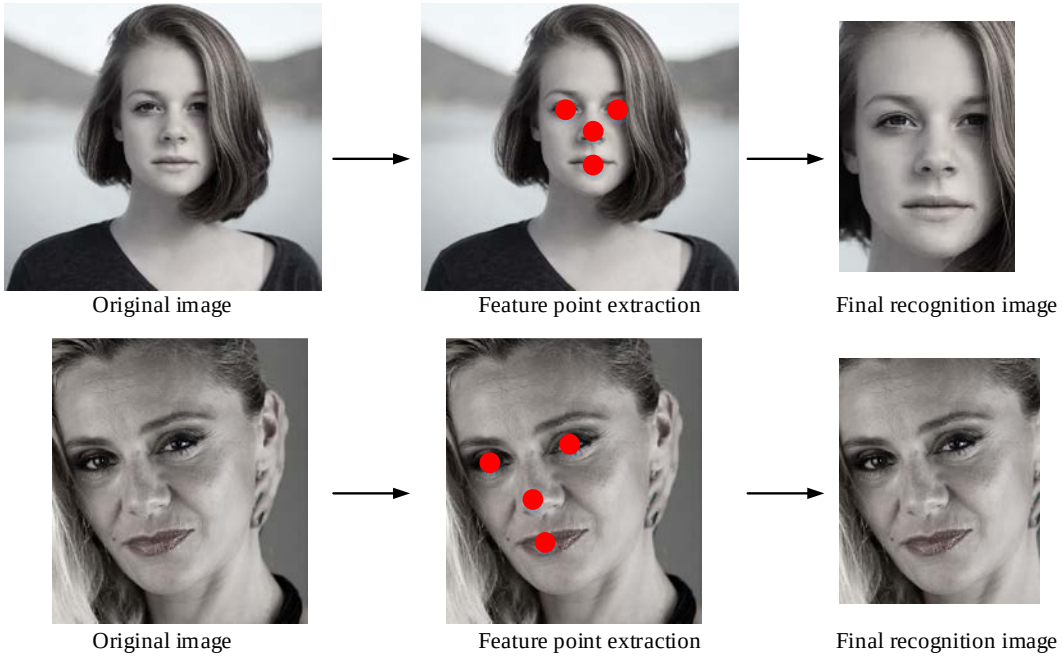


Fig. 15. Face recognition process of different angles and image scales.

It can be seen that the model realizes the frame selection of face scale frame by extracting key feature points. The specific experimental data are shown in **Table 6**.

Table 6. Recognition accuracy of different models in different data sets

method	CK+	Oulu-CASIA	JAFFE	CARE
STGCN+AM+PF	0.973	0.905	0.954	0.602
R. Zhao et al：GCN	0.978	0.881	-	0.546
B. Yang et al：CNN	0.970	0.923	0.922	-

In **Table 6** above, in CK+ data set, there is little difference in the recognition accuracy of each model, which is within the range of (0.97,0.98). However, in the Oulu-CASIA dataset, the difference between the models is too large. The recognition accuracy of the STGCN+AM+PF model used in the study reaches 90.53%, while the recognition accuracy of the GCN model is the worst, which decreases by 2.43% compared with the STGCN+AM+PF model. In the small JAFFE data set, the performance of STGCN+AM+PF model is improved, and the recognition accuracy is increased by 3.2% compared with the CNN model. In a large CARE dataset, the performance of both models decreased significantly, but the recognition accuracy of STGCN+AM+PF model was still 5.6% higher than that of GCN model. To sum up, the performance of the face recognition model designed in the study is among the best in most data sets.

To further verify the recognition accuracy of this model for expressions that are easily confused, the DFEPN dataset was introduced in this study, and the three expressions of calm, crying and pain were recognized in it. The experimental results were compared with the CNN model proposed by B. Yang et al., as shown in **Fig. 16**.

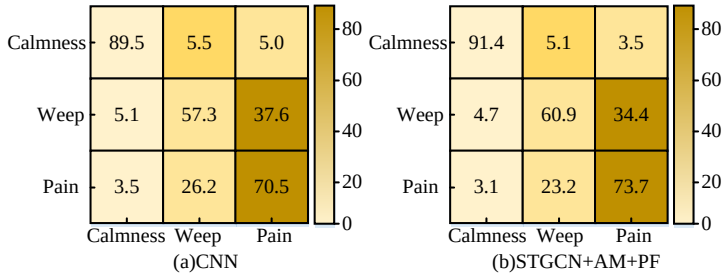


Fig. 16. Performance matrix comparison of different models in confusing expressions

Fig. 16 shows that the GNN model and the research adoption model have relatively accurate recognition accuracy for calm expression, with 89.5% and 91.4% respectively. Overall, the research adoption model has a slightly higher accuracy rate of 1.9%. However, both models have a high confusion rate for crying and pain expressions. The GNN model's accuracy rate for recognizing crying expressions is 57.3%, which is lower than the model used in the study by 3.6%. Additionally, the model has a confusion rate of 37.6% and 34.4% for pain expressions, respectively. The study found that the model used had a 3.2% higher accuracy in recognizing painful expressions compared to the GNN model. **Table 6** shows that the performance gap between the two models is not significant, and even compares their advantages and disadvantages in some datasets. However, when it comes to easily confused expressions, the model used in the study has higher recognition accuracy than the GNN model with similar performance. Therefore, the spatio-temporal graph convolutional network model, based on a sliding block self-attention mechanism and peak frame, can achieve better expression recognition. To evaluate the performance of the enhanced model, a comparative experiment was conducted on a randomly selected dataset. The average facial recognition results of CNN, LSTM, STGCN, STGCN-SA and the improved model are shown in **Fig. 17**.

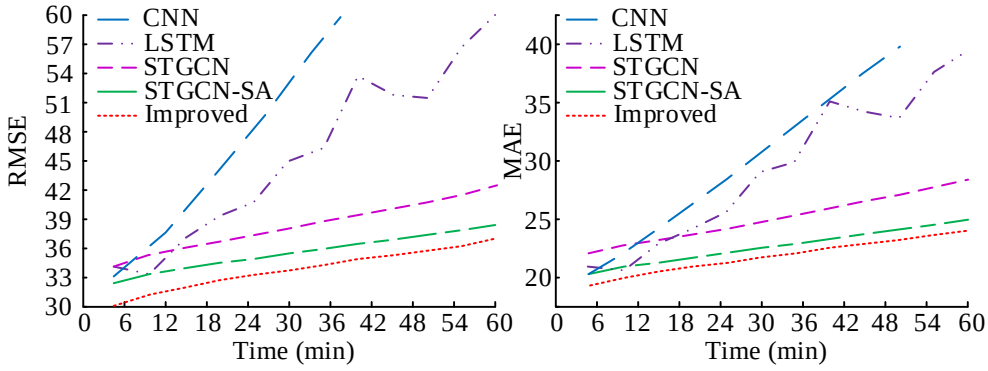


Fig. 17. The average result of facial recognition by the system

Fig. 17(a) shows that the RMSE of the five models studied gradually increases over time. The improved model has an RMSE of 32.82, an STGCN of 38.29, and an STGCN-SA of 35.64. In comparison, the improved model has lower recognition errors. **Fig. 17(b)** indicates that the improved model has the lowest MAE, which is 21.80, suggesting that its recognition results are more reliable than those of other models. **Fig. 18** displays the configuration interface of the facial recognition system designed for research.

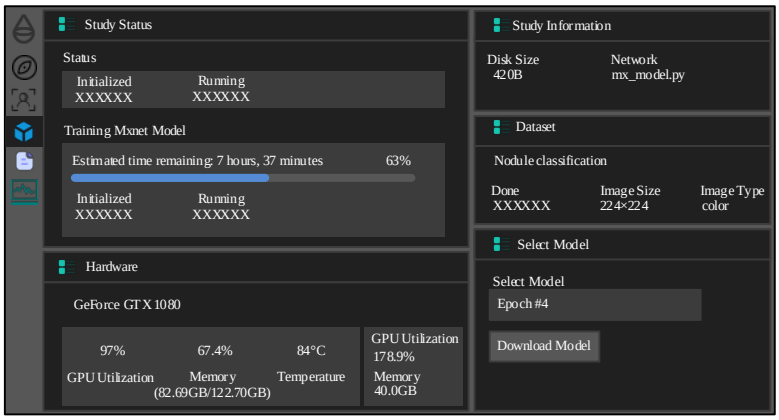


Fig. 18. System configuration interface

5. Discussion

In the application of intelligent libraries, it can be seen that the spatio-temporal convolutional network model based on the sliding block self-attention mechanism and the peak frame image has relatively good expression recognition ability. This is because the spatio-temporal graph convolutional network selected in the study can take care of the feature extraction of time and space dimensions at the same time, and its learning ability also further improves the effect of deep learning in complex data. Among them, graph is a technology that can extract the structural relationship between data, which often can get more key information than the traditional one by one analysis method. However, the underlying pattern of graph structure is more complex, so dimensionality reduction is a necessary optimization process of graph structure. This paper introduces the sliding block self-attention mechanism, which can improve the generalization of the model, strengthen the uniqueness of the feature data, and increase the model's attention to local areas, so as to help achieve better image classification effect. In order to further enhance the extraction effect of texture features, the theory of peak frame image is introduced, and the suitable feature fusion method is used to combine the peak frame image with the spatio-temporal map features. Finally achieve better expression recognition effect.

6. Conclusion

Under the empowerment of technology, libraries are constantly upgrading and transforming towards intelligence. The research proposes an adaptive mechanism model combining peak frame images and spatio-temporal maps, which can handle the spatio-temporal feature information of dynamic images, but also solve the problems of weak generalization ability of the model and poor texture feature recognition. This study utilises the CK+ and Oulu-CASIA datasets to simulate the enhanced model. This paper conducts a comparative analysis of the three edge connection modes of the spatio-temporal graph convolution network model. The results indicated that the full connection method is capable of achieving the most favourable recognition effect, leading to an improved accuracy of 1% to 2.15% as compared to other connection methods. This demonstrated that the full-connection method can extract dynamic expression features more effectively. The analysis of introduced adaptive modules yielded the optimal number of such modules. The

usage of one adaptive module in the CK+ dataset and four adaptive modules in the Oulu-CASIA dataset resulted in improved recognition accuracy of 0.88% and 2.31%, respectively. This illustrated the usefulness of the adaptive mechanism in enhancing recognition accuracy of the model. To further extract the image's texture and other details, the peak frame image was integrated into the original model, and the two fusion methods of the feature layer and the decision layer are compared. The results demonstrated that the decision level fusion method achieved higher performance, improving the recognition accuracy by 3.35% to 11.8%, compared to that of the spatio-temporal graph convolutional network model. Compared to other recognition models, the STGCN+AM+PF model achieved a 3.2% improvement in recognition accuracy on the small JAFFE dataset, surpassing the performance of the CNN model. In large CARE datasets, the recognition accuracy of the STGCN+AM+PF model remained 5.6% higher than that of the GCN model. Furthermore, the matrix analysis revealed that the confusion rate of the STGCN+AM+PF model was 3.2% lower than that of the GCN model when confronted with expressions of pain and crying that are often confused. This technology can improve the accuracy of facial recognition and is of great significance for the intelligent development of libraries. However, due to the limitations of academic knowledge, there is still some room for improvement. As the method is heavily influenced by the key point annotation, the model becomes less effective when the images are occluded, etc. Therefore, subsequent research should focus on the recognition of missing images to improve the applicability of the model.

Acknowledgment

The research is supported by A 2021 study of Culture and Tourism in Shandong Province: “Research in the excavation and utilization of the red literatue in Jiaodong district library” (No, 21 WL (H) 17).

The research and application of facial expression recognition systems are conducted at the Yantai Vocational College Library.

References

[1] Z. Hao, X. Wang, and S. Zheng, “Recognition of basketball players' action detection based on visual image and Harris corner extraction algorithm,” *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, vol.40, no.4, pp.7589-7599, Aug. 2021. Article (CrossRef Link)

[2] S. A. Rizwan, Y. Y. Ghadi, A. Jalal, and K. Kim, “Automated Facial Expression Recognition and Age Estimation Using Deep Learning,” *Computers, Materials & Continua*, vol.71, no.3, pp.5235-5252, Jan. 2022. Article (CrossRef Link)

[3] Z. Wang, J. Zhan, C. Duan, X. Guan, and K. Yang, “Vehicle detection in severe weather based on pseudo-visual search and HOG–LBP feature fusion,” *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol.236, no.7, pp.1607-1618, Jun. 2022. Article (CrossRef Link)

[4] R. Mu and X. Zeng, "A Review of Deep Learning Research," *KSII Transactions on Internet and Information Systems*, vol. 13, no. 4, pp. 1738-1764, 2019. Article (CrossRef Link)

[5] D. Lu, D. Wang, K. Zhang, and X. Zeng, “Age estimation from facial images based on Gabor feature fusion and the CIASO-SA algorithm,” *CAAI Transactions on Intelligence Technology*, vol.8, no.2, pp.518-531, Jun. 2023. Article (CrossRef Link)

[6] Z. Sun, R. Chiong, and Z.-P. Hu, "Self-adaptive feature learning based on a priori knowledge for facial expression recognition," *Knowledge-Based Systems*, vol.204, no.1, Sep. 2020. Article (CrossRef Link)
[7] M. T. B. Iqbal, B. Ryu, A. R. Rivera, F. Makhmudkhujaev, O. Chae, and S.-H. Bae, "Facial Expression Recognition with Active Local Shape Pattern and Learned-Size Block Representations," *IEEE Transactions on Affective Computing*, vol.13, no.3, pp.1322-1336, Jul.-Sept. 2022. Article (CrossRef Link)
[8] C. Liu, K. Hirota, J. Ma, Z. Jia, and Y. Dai, "Facial Expression Recognition Using Hybrid Features of Pixel and Geometry," *IEEE Access*, vol.9, no.1, pp.18876-18889, Jan. 2021. Article (CrossRef Link)
[9] S. D. Rajagopal, and B. Ramachandran, "3D face expression recognition with ensemble deep learning exploring congruent features among expressions," *Computational Intelligence*, vol.38, no.2, pp.345-365, Apr. 2022. Article (CrossRef Link)
[10] K. V. Swaroop and M. S. Saravanan, "Prediction of Human Facial Expression with Machine Learning Classifier Using Convolutional Neural Network Instead of Traditional Pixel Value Algorithm for Better Accuracy," *ECS Transactions*, vol.107, no.1, pp.13937-13949, Apr. 2022. Article (CrossRef Link)
[11] S.A.M. Al-Sumaidaee, M.A.M. Abdullah, R.R.O. Al-Nima, S.S. Dlay, and J.A. Chambers, "Spatio-temporal modelling with multi-gradient features and elongated quinary pattern descriptor for dynamic facial expression recognition," *Pattern Recognition*, vol.142, Oct. 2023. Article (CrossRef Link)
[12] H. Zaaraoui, S. E. Kaddouhi, M. Abarkan, "A novel face recognition approach based on strings of minimum values and several distance metrics," *International Journal of Computer Aided Engineering and Technology (IJCAET)*, vol.18, no.1/2/3, pp.60-76, Nov. 2023. Article (CrossRef Link)
[13] R. Sudharsanan, P.V. Gopirajan, and K. S. Kumar, "Efficient Feature Extraction from Multispectral Images for Face Recognition Applications: A Deep Learning Approach," *Journal of Physics: Conference Series*, vol.1767, no.1, Feb. 2021. Article (CrossRef Link)
[14] D. Shirafuji, R. Rzepka, and K. Araki, "Argument Extraction for Key Point Generation Using MMR-Based Methods," *IEEE Access*, vol.9, no.1, pp.103091-103109, Jul. 2021. Article (CrossRef Link)
[15] S. Wenshun, S. Yanwen and X. Liujing, "Research on will-dimension SIFT algorithms for multi-attitude face recognition," *High Technology Letters (English Version)*, vol.28, no.3, pp.280-287, Sep. 2022. Article (CrossRef Link)
[16] J. Liu and Y. Che, "Action recognition for sports video analysis using part-attention spatio-temporal graph convolutional network," *Journal of Electronic Imaging*, vol.30, no.3, Jun. 2021. Article (CrossRef Link)
[17] D. Zhang, Y. Peng, Y. Zhang, D. Wu, H. Wang, and H. Zhang, "Train Time Delay Prediction for High-Speed Train Dispatching Based on Spatio-Temporal Graph Convolutional Network," *IEEE Transactions on Intelligent Transportation Systems*, vol.23, no.3, pp.2434-2444, Mar. 2022. Article (CrossRef Link)
[18] X. Wang, M. Cheng, J. Eaton, C.-J. Hsieh, and S. Felix Wu, "Fake Node Attacks on Graph Convolutional Networks," *Journal of Computational and Cognitive Engineering*, vol.1, no.4, pp.165-173, Nov. 2022. Article (CrossRef Link)
[19] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *Computational Visual Media*, vol.8, no.3, pp.331-368, Sep. 2022. Article (CrossRef Link)
[20] H. Wei, W. Zhou, X. Zhou, and Z. Duan, "Sub-Frame Layer Rate Distortion Optimization for High Efficiency Video Coding," *IEEE Access*, vol.8, pp.53116-53132, Mar. 2020. Article (CrossRef Link)
[21] X. Ai, M. Sheng, X. Su, S. Ai, X. Jiang, S. Yang, and Y. Ai, "Effects of frame beam on structural characteristics of artificial soil on railway cut-slopes in southwestern China," *Land Degradation & Development*, vol.32, no.1, pp.482-493, Jan. 2021. Article (CrossRef Link)

[22] J. Bai, J. Zhu, Y. Song, L. Zhao, Z. Hou, R. Du, and H. Li, "A3T-GCN: Attention Temporal Graph Convolutional Network for Traffic Forecasting," *ISPRS International Journal of Geo-Information*, vol.10, no.7, Jul. 2021. Article (CrossRef Link)
[23] Y. Liu, C. Feng, X. Yuan, L. Zhou, W. Wang, J. Qin, Z. Luo, "Clip-aware expressive feature learning for video-based facial expression recognition," *Information Sciences*, vol.598, pp.182-195, Jun. 2022. Article (CrossRef Link)
[24] T. Ma, W. Tian, Y. Xie, "Multi-level knowledge distillation for low-resolution object detection and facial expression recognition," *Knowledge-Based Systems*, vol.240, Mar. 2022. Article (CrossRef Link)
[25] F. Nan, W. Jing, F. Tian, J. Zhang, K.-M. Chao, Z. Hong, Q. Zheng, "Feature super-resolution based Facial Expression Recognition for multi-scale low-resolution images," *Knowledge-Based Systems*, vol.236, Jan. 2022. Article (CrossRef Link)
[26] N. B. Kar, D. R. Nayak, K. S. Babu, Y.-D. Zhang, "A hybrid feature descriptor with Jaya optimised least squares SVM for facial expression recognition," *IET Image Processing*, vol.15, no.7, pp.1471-1483, May. 2021. Article (CrossRef Link)
[27] W. Zhang, X. Zhang, Y. Tang, "Facial expression recognition based on improved residual network," *IET image processing*, vol.17, no.7, pp.2005-2014, May 2023. Article (CrossRef Link)
[28] Z. Shang, H. Xie, Z. Zha, L. Yu, Y. Li, and Y. Zhang, "PRRNet: Pixel-Region relation network for face forgery detection," *Pattern Recognition*, vol.116, Aug. 2021. Article (CrossRef Link)
[29] B. Yang, J. Cao, R. Ni, and Y. Zhang, "Facial Expression Recognition Using Weighted Mixture Deep Neural Network Based on Double-Channel Facial Images," *IEEE Access*, vol.6, pp.4630-4640, Dec. 2018. Article (CrossRef Link)
[30] R. Zhao, T. Liu, Z. Huang, D. P.K. Lun, and K.-M. Lam, "Spatial-Temporal Graphs Plus Transformers for Geometry-Guided Facial Expression Recognition," *IEEE Transactions on Affective Computing*, vol.14, no.4, pp.2751-2767, Oct.-Dec. 2023. Article (CrossRef Link)

Yan Qu was born in Zhengzhou city, He Nan province, HN, CHN in 1976. She received the B.S. in Economics from Beijing Institute of Technology, Beijing, China in 2004, and she received the M.S. in Business Administration from Tianjin University, Tianjin, China in 2005.
From 2018 to 2024, she was a librarian in Yantai Vocational College Library. She has been an associate researcher since 2020. She is the moderator of the two topics of Yantai city and Shang Dong province, she is the participant in more than five project researches. She is the author of more than 7 articles. Her research interests include library management, information resource in library, reading promotion, construction of intellectual library, literature data mining.

Yan Liu was born in Qingdao city, Shan Dong province, SD, CHN in 1971. She received the B.S. in Archival Secretarial from the party school of CPC Shandong Provincial Committee, Shandong, China in 2004.
From 2005 to 2024, she was a librarian in Yantai Vocational College Library. She is the author more than 11 articles. Two of the articles has won the prize of excellence by Shandong Library Association. Her research interests include library management, information resource in library, reader service.