Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **10**, Number **1**, February **2025**

Deep Learning Model-based Emotion Recognition for Visual Communication in Human-computer Interaction

Zhen-Bin Huang*
Sanda University, Shanghai 201209, P. R. China
leohuang98@126.com

Xin-Yu Chen
Sanda University, Shanghai 201209, P. R. China
chenxinyuuk@163.com

Evan Huang
College of Engineering and Information Technology
Cavite State University, Lahug, Cebu 6000, Philippines
ua4706@163.com

*Corresponding author: Zhen-Bin Huang
Received March 5, 2024, revised June 26, 2024, accepted August 2, 2024.

Abstract. *Visual emotion recognition is of great significance in economic, social, and scientific technology, and has a wide application market in the field of human-computer interaction. In the process of human-computer interaction, accurate understanding of the user's emotional state is crucial for improving user experience, personalized recommendation and sentiment analysis. By applying visual emotion recognition technology, computers can recognize the user's emotional state by analyzing the user's facial expression, posture, and other visual information, so as to interact and respond to the user more effectively. With the rapid development of artificial intelligence technology, emotion recognition of visual information using deep learning models plays an increasingly important role in human-computer interaction. In order to improve the accuracy of visual emotion recognition, this paper proposes a visual emotion recognition method that integrates self-supervised learning and attention mechanism. The image semantic information is learned through a self-supervised auxiliary task assisted by a trunk feature extraction network, and the emotion mapping map is generated based on the attention mechanism to enhance the feature representation. Experimental validation is carried out on several emotion datasets, and the results show that compared to the method without self-supervised learning, the proposed method significantly improves the emotion recognition accuracy and verifies its effectiveness in practice. The method is important for enhancing the accuracy of visual emotion recognition, improving the human-computer interaction experience, as well as enhancing the intelligent interaction capability of the system.*
Keywords: Visual emotion recognition; Deep learning; attention mechanisms; Self-supervised learning

1. **Introduction.** With the continuous development and popularization of intelligent technology, research and applications in the field of human-computer interaction have received increasing attention. Understanding and recognizing the user's emotional state becomes crucial in the interaction process between humans and computers. Visual emotion

recognition, as an important technical tool [1, 2], can help computer systems understand the user's emotional feedback, so as to better satisfy the user's needs and enhance the user experience. In traditional human-computer interaction, it often relies on the user's text input or simple interaction commands for response and interaction. However, this approach often fails to fully capture and understand the user's true emotional state [3]. With the continuous development and application of deep learning models, emotion recognition technology based on visual information is gradually becoming a research hotspot in the field of human-computer interaction [4, 5]. By analyzing the user's facial expression, posture and other visual information, the deep learning model can effectively identify and understand the user's emotional state, thus realizing a more intelligent and personalized interaction experience.

In previous studies, research scholars analyzed visual information for emotion based on some manual features, but the discriminative ability of manual features is limited, with the massive growth of data and the rapid development of deep learning, research scholars began to study visual emotion based on deep learning. Emotion as an abstract concept, understanding the semantic information in the picture helps in emotion classification [6, 7]. However, semantic annotation of the dataset is costly, the semantic richness of the features extracted only by image-level emotion labeling is limited, and different regions of the image have different effects on emotion. To address these issues, this paper proposes a visual emotion recognition method that incorporates self-supervised learning and attention mechanisms. By adding a self-supervised auxiliary task of picture rotation, the backbone feature extraction network is able to extract richer semantic features, such as the type, position, and pose of the objects in the pictures, which reduces the labeling burden. By simulating the human visual processing mechanism, we generate emotion mapping maps based on spatial attention and channel attention in the emotion classification task, and strengthen the original features by coupling operations to make them more discriminative. Experimental results show that the model has improved sentiment recognition accuracy on multiple datasets compared to the method without self-supervised learning.

The structure of this paper is as follows: Firstly, it introduces the related research results and the current status, secondly, it gives a brief introduction to the related theoretical foundation, then it describes the method of emotion recognition for visual communication based on deep learning, completes the experiments of the method, and finally the method of this paper is summarized.

1.1. **Related work.** With the development of neural network research, deep learning methods have achieved great success in computer vision. Rao et al. [8] combined image aesthetics and the overall and local low-level visual features of an image, and proposed a deep network incorporating different output layers of convolutional neural networks, which learns multilevel deep feature representations used for image emotion recognition, and efficiently categorizes the emotions of different types of images such as scenery, people, and so on. Song et al. [9] argued that the image emotions may be generated by certain spatial regions only They proposed an emotion network with visual attention, they used a multilayer network to generate attention distribution maps of image regions, and integrated the visual attention framework with the convolutional neural network emotion classification framework, trained SentiNet-A in an end-to-end manner, and achieved better classification results on multiple datasets. She et al. [10] proposed a weakly supervised network, in which a weakly supervised mechanism is introduced in emotion recognition, and only the convolutional neural network is utilized for emotion classification. A weakly supervised mechanism is introduced in the emotion recognition to generate the emotion

stimulus map using only global image-level labels, which reduces the label burden. They used a cross-space pooling strategy to train a convolutional neural network to generate stimulus maps corresponding to a specific emotion in the emotion stimulus region detection branch, and coupled the detected emotion stimulus maps with deep emotion features as feature vectors in the emotion classification branch.

Yang et al. [11] designed an algorithm to automatically discover emotion regions in images by first generating candidate proposals using an existing object detection tool and removing the redundant and noisy proposals, and then connect a convolutional neural network to each candidate proposal to compute the sentiment score, and automatically discover emotional regions under the multiple factors of object score and sentiment score. The unsupervised learning approach can significantly reduce the labeling burden. In the next visual emotion research, more consideration should be given to unsupervised learning of emotion-related information in images, extracting more discriminative features, and improving model accuracy. Li et al. [12] assign an emotion value to each ANP concept in SentiBank, detect ANPs in the picture and get the corresponding emotion value of the picture by calculation, train a one-dimensional logistic regression classifier In order to predict the sentiment of a picture based on its sentiment value, the ANPs detected in the picture are used as intermediate representations to train a regularized logistic regression classifier, which fuses the two kinds of sentiment prediction outputs and further improves the performance of the network. Yamamoto et al. [13] proposed an image sentiment recognition method that uses both semantic and visual features. Semantic features are created by object detection methods and word embedding models, semantic features are concatenated with visual features generated by a fine-tuned convolutional neural network, and sentiment classification is performed through the concatenated feature vectors. Zhang et al. [14] argued that the semantic information of different objects in a picture and their relationship with each other are closely related to sentiment, and proposed a graph-based object semantic refinement model to extract visual sentiment classification for multilevel semantic features. Where the graph structure is used to represent the object semantic information and its positional relationship in the image, and the refined semantic information is combined with the overall visual features of the picture to achieve more accurate emotion recognition. However, emotion is not only related to semantic information, there are many hidden visual information that play a key role in emotion recognition. Some researchers and scholars have found that emotion polarity is related to low-level features in convolutional neural networks, and specific emotion categories are related to high-level features in convolutional neural networks. Therefore, the fusion of various emotion-related information should also be considered in subsequent research.

1.2. **Motivation and contribution.** From the current state of research, it can be seen that the research work on visual emotion recognition has made great progress. But the connection between semantic information in pictures and emotion is very close, however, the existing public dataset lacks semantic labels, and the semantic information in the features extracted by deep learning networks is not rich enough, which makes it difficult to effectively utilize the close connection between emotion and the semantic information in the pictures, and leads to the low accuracy of model recognition. The main contributions of this paper, which focuses on the current problems in visual emotion recognition, are as follows:

A visual emotion recognition method that integrates self-supervised learning and attention mechanisms. Aiming at the problem of limited semantic information richness in the features extracted by deep learning networks, this paper considers learning the semantic information in the image, such as the type, position, and pose of the object, through

self-supervised auxiliary tasks to increase the semantic richness of the features, and investigates the reinforcing effect of the attention mechanism on the features to improve the accuracy of the model emotion recognition.

2. **Relevant theoretical analysis.**

2.1. **Convolutional neural network.** Deep learning is the current mainstream research direction of machine learning. The concept of deep learning comes from researchers' study of artificial neural networks, in which a multilayer perceptron containing multiple intermediate hidden layers is a typical deep learning structure [15]. Researchers hope that deep learning will enable computers to mimic the mechanisms of the human brain and build neural network systems that can learn and analyze on their own, and can interpret and learn from data such as images, sounds, and texts just like humans. Early datasets are small, researchers can use support vector machines, random forests, and other machine learning algorithms to achieve good recognition accuracy on these small datasets. With the rapid popularization of the Internet, image data is becoming more and more easily accessible, and researchers have collected larger datasets, such as ImageNet, LabelMe, etc. To process these datasets, researchers need a model with more powerful learning ability, and deep learning has come into the researchers' sight [16, 17].

In 2012, a researcher proposed the AlexNet model based on a convolutional neural network [18], and the recognition accuracy on ImageNet is better than the previous algorithms. The AlexNet model mainly consists of five convolutional layers and three fully connected layers, and the whole model contains 60 million parameters and 650,000 neurons. To solve the overfitting problem, Alex uses Dropout regularization. The success of AlexNet brought a huge change to the academic world at that time and set off a deep learning boom. In 2014, a researcher found that two 3×3 convolutional layers in series had the same large receptive field as a 5×5 convolutional layer, and three 3×3 convolutional layers in series had the same receptive field as a 7×7 convolutional layer. The VGG model is proposed on this basis. By using 3×3 convolutional layers in the network and increasing the depth of the network, i.e., reducing the number of parameters and making the network more capable of learning features. In 2014, Google researchers proposed the GoogleNet model. In this model, the Inception Module is used, multiple branches are used to extract higher-order features at different scales, the depth and width of the network are increased while the network parameters are reduced, and a 1×1 convolutional layer is used in the last layer of the network in place of the fully connected layer. Theoretically, the accuracy of the deep learning model should become better as the depth of the network deepens, but there is a degradation problem for deep networks, and the effect of the network is worse beyond a certain depth. In 2015, the ResNet model was proposed, which breaks the limitation of the depth of the network to a certain extent. A constant mapping is added to the ResNet model, as shown in Figure 1, the original network needs to learn the function $F(x)$, after adding the mapping, the function that needs to be learned is $F(x) + x$. When the network is backpropagated, the gradient of the next layer can be directly passed to the network of the previous layer, which solves the problem of gradient disappearance caused by the depth of neural networks [19]. The depth of the ResNet network they designed can reach more than one hundred layers, and in this study, ResNet-101 was used as the feature extraction network for images.

2.2. **Self-supervised Learning.** In recent years, the field of computer vision has developed rapidly with the application of deep learning techniques. When performing tasks

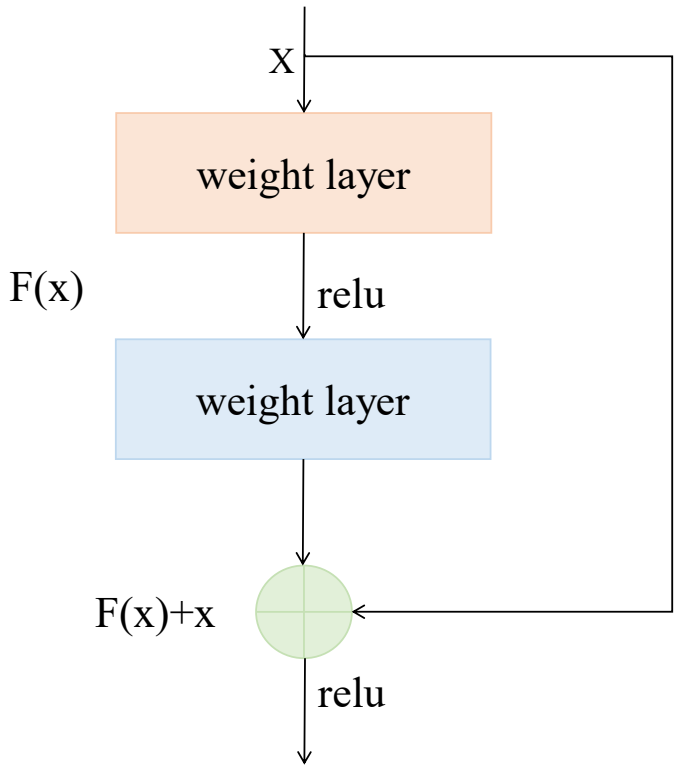

Figure 1. ResNet residual structure

such as classification, recognition, and detection, neural networks can extract visual features that are discriminative to the task from labeled large datasets. Supervised learning-based methods require a large number of manually labeled labels, but manual labeling of datasets is time-consuming and laborious. Self-supervised learning is a kind of unsupervised learning, which can train neural networks without manual labeling, and is one of the research hotspots of deep learning in recent years [20, 21, 22]. Self-supervised learning is customized auxiliary tasks, according to which supervisory information is mined from a large amount of unlabeled data to provide supervisory signals for network training. The methods of self-supervised learning can be categorized into context-based self-supervised learning, timing-based self-supervised learning, and contrast-constraint-based self-supervised learning. For example, predicting the color of a picture, the relative position of each block of a picture and the relative position of objects in each frame of a video. Some researchers and scholars have used self-supervised learning in other domains [23, 24], which is applicable to various downstream tasks by solving customized auxiliary tasks, where trained neural network models can learn more robust and broader features from unlabeled data rather than features specific to a particular type of task.

In the visual emotion recognition task, this paper considers a multi-task setup where a backbone feature extraction network is trained to be used on both the main emotion classification task and the self-supervised auxiliary task.

Unlike most multitasking setups that aim to achieve the best accuracy on all tasks simultaneously, in this paper, we only aim to improve the emotion recognition accuracy of the main task through the self-supervised auxiliary task. Some researchers define the auxiliary task as predicting the rotation angle of an image, and automatically learn image features by training a neural network to accomplish the auxiliary task. Each image in the dataset is rotated by 0°, 90°, 180°, and 270° to obtain a new dataset, and the rotation angle of each image is defined as an image label. This task may seem simple, but actually

provides a very powerful supervised signal for semantic feature learning. To make correct rotation angle predictions, the neural network learns for the objects in the image and their locations. As a result, the features extracted by the neural network contain rich semantic and positional localization information.

In the research of visual emotion recognition, emotion is closely linked to the semantic information in the pictures. By learning the semantic information of the pictures, it can help the model to judge the emotion of the pictures. However, it is time-consuming and laborious to annotate the detailed semantic information of pictures, and this paper hopes to learn the semantic information of pictures in an unsupervised way. By predicting the rotation angle of a picture, the network can automatically learn the semantic features such as object type, position and pose in the picture, which are very helpful for the study of the emotion of the picture. Therefore, this paper investigates the addition of picture rotation self-supervised auxiliary task in emotion recognition network to improve the model emotion recognition performance.

2.3. **Attention mechanism.** In the learning process of neural network model, in order to enhance the expression ability of the model and better learn the data features, usually the model is more complex, the number of model parameters is larger, which will lead to an information overload, sometimes affecting the learning effect of the model. In order to solve the problems caused by this situation, the use of attention mechanisms to focus on the information learned from the data, extract the more critical information under the current task, reduce other data that have an impact on the model learning effect, so the model can have a higher processing efficiency. Because the attention modeling mechanism is effective in extracting features and also has plug-and-play characteristics, the attention mechanism is considered to be introduced in the model.

The attention mechanism is mainly divided into three steps. First of all, it is necessary to score the input data with attention as shown in Equation (1), and the specific scoring functions mainly used are additive function, dot product operation function, scaling dot product function, etc. The specific scoring function has an additive function, dot product operation function, scaling dot product function, etc., that are mainly used.

$$x_i = F(x, w, k) \tag{1}$$

The scored data is then numerically transformed to obtain the weight coefficients required for the attention mechanism, where softmax is commonly used, as shown in Equation (2).

$$\hat{o}_i = \frac{\exp(x_i)}{\sum_{i=1}^{N} \exp(x_i)} \tag{2}$$

Finally, the weight coefficients are weighted and summed over the data to highlight important data features that can be used for subsequent training as shown in Equation (3).

$$s = \sum_{j=1}^{N} \hat{o}_j x \tag{3}$$

3. Visual emotion recognition incorporating self-supervised learning and attention mechanisms.

3.1. **Modeling framework.** A visual emotion recognition network framework incorporating self-supervised learning and attention mechanisms proposed in this paper is shown in Figure 2.

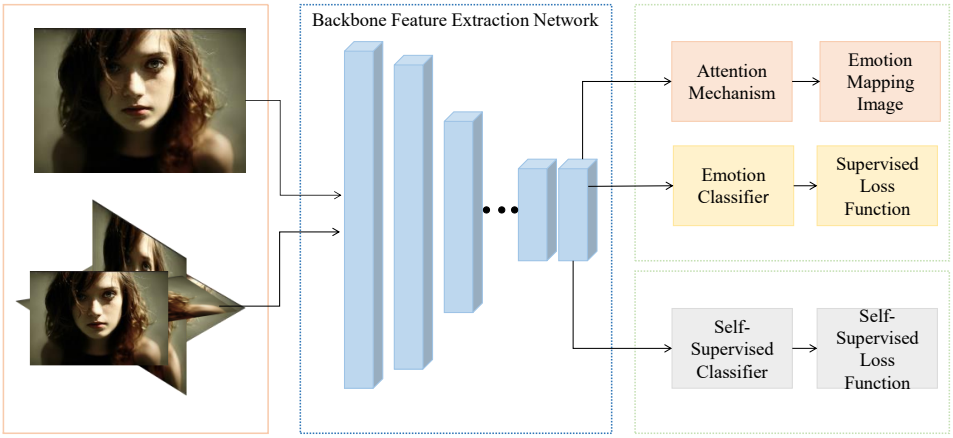

Figure 2. Visual emotion recognition network framework diagram

The model integrates the emotion classification task and the picture rotation task, through the learning of the picture rotation angle, which enables the feature extraction network to learn the rich semantic information such as the object type, position, and pose in the picture, and get the corresponding emotion mapping map through the attention mechanism to locate the emotion stimulation region, and improve the discriminative ability of the features.

The whole model can be divided into three parts: the backbone feature extraction network, the picture rotation self-supervised auxiliary module, and the emotion classification module based on the attention mechanism. Firstly, the picture is preprocessed and the rotated picture is obtained by flip and transpose operations, which is input into the backbone feature extraction network, and then the extracted features are input into the emotion classification module and self-supervised module respectively to predict the picture emotion and rotation angle.

In this paper, ResNet-101 is used as the basic model for extracting picture features. The last two layers (global average pooling layer and fully connected layer) of ResNet-101 are removed to obtain the backbone feature extraction network. Next, the self-supervised auxiliary module for picture rotation and the sentiment classification module based on the attention mechanism are introduced.

3.1.1. *Image rotation self-supervision assist module.* We hope that an unsupervised learning approach makes the convolutional neural network $F(\cdot)$ able to learn some rich semantic information, such as the type and relative position of the objects in the picture, which are very helpful for emotion recognition. In order to realize such a goal, this paper designs the rotary self-supervised auxiliary module to predict the geometric transformations of pictures by training the neural network model $F(\cdot)$.

Define a set $G = \{g(\cdot|k)\}_{k=1}^{k}$ containing k geometric transformations, where $g(\cdot|k)$ denotes the geometric transformation of the picture X to produce the picture $X^k = g(x|k)$ with label k. Use the picture X^k as the input to the model $F(\cdot)$, k^* as the label of the picture, and the output as the probability distribution of all possible geometric transformations as shown in Equation (4):

$$F(X^k|\theta) = \{F^k(X^k|\theta)\}_{k=1}^{k} \tag{4}$$

where $F^k(X^k|\theta)$ denotes the prediction rate of a geometric transformation of a picture into a label k, and θ is a learnable parameter for model $F(\cdot)$. Thus, given a dataset $D = \{X_i\}_{i=1}^N$ containing N images, the problem to be solved by the self-supervised learning model $F(\cdot)$ is:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} Loss(X_i, \theta) \tag{5}$$

where $Loss(\cdot)$ is defined as:

$$Loss(X_i, \theta) = -\frac{1}{K} \sum_{K=1}^{K} \log(F^K(g(X_i|k)|\theta)) \tag{6}$$

The prediction of geometric transformations of pictures allows the model to learn features that are useful for a number of visual recognition tasks. For example, in order to successfully predict the rotation angle of a picture, the model must learn to understand the scene depicted in the picture, learn to localize the more obvious objects in the picture, understand the relationships between the objects, and so on. This semantic information is useful for understanding picture emotions. Therefore, in this paper, we define the self-supervised task as predicting the rotation angle of a picture by rotating the input picture by a number of angles respectively, and predicting it, so that the features extracted by the model $F(\cdot)$ are rich in semantic information.

Define operator A as a clockwise rotation B degrees of the picture X. By processing the pictures in the dataset, a dataset C containing R rotational variations is obtained.

$$g(X|r) = Rot(X, (r-1) * \phi) \tag{7}$$

For example, after rotating a picture clockwise by 0, 90, 180, and 270 degrees, a new dataset with four rotations is obtained $G = \{g(X|r)\}_{r=1}^4$. The process of the picture is as follows: the original picture is rotated clockwise by 0 degrees; for 90-degree rotation, the picture is first transposed and then flipped vertically; for 180-degree rotation, the picture is first flipped vertically and then horizontally; for 270-degree rotation, first flip the picture vertically and then transpose it.

The pictures in the rotated expanded dataset G are processed by the feature extraction network to obtain the feature f_{rot}. f_{rot} is input into the rotated self-supervised classifier to predict the rotation angle of the pictures, and the network framework of the rotated self-supervised classifier is shown in Figure 3. The feature vector d_{rot} is generated by passing the feature f_{rot} sequentially through the global average pooling layer and the fully connected layer, and the predicted probability of the rotation angle of the picture is calculated by Softmax function. The loss function for this task is shown in Equation (8):

$$L_{rot}(\theta, r; X) = -\frac{1}{N_{rot}} \sum_{i=1}^{N_{rot}} \sum_{r=1}^{R} 1(z_i = r) \log(F^r(X_i|\theta)) \tag{8}$$

where N_{rot} denotes the total number of pictures in the dataset G after rotational transformation, R denotes the number of rotational changes performed on each picture, z_i denotes the rotation angle label of the picture X_i, and $F^r(X_i|\theta)$ denotes the probability that the ith picture in the dataset G is predicted to be rotated to a rotation angle $r(s) = 1$ when the condition s is true, and 0 otherwise.

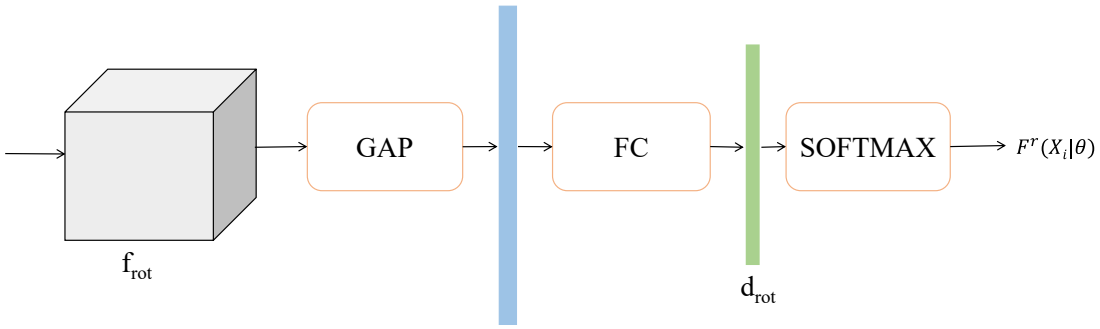

Figure 3. Structure of rotary self-supervised classifier

3.1.2. *Sentiment classification module based on attention mechanism.* It has been found that different parts of a picture elicit emotions differently. Inspired by the structure of CBAM, this paper adds spatial attention and channel attention to the emotion categorization task to generate emotion mapping maps and reinforce features. In this chapter, channel attention and spatial attention are processed on input features in parallel. Firstly, feature f^1 is obtained by the backbone feature extraction network, and a 1×1 convolution operation is performed on it to compress the spatial dimension to obtain feature f^2. Then, a concatenation operation of spatial and channel attention is performed on it to obtain the reinforced feature f^3, as shown in Equation (9):

$$f^3 = f^2 \otimes M_c(f^2) \otimes M_s(f^2) \tag{9}$$

where $\otimes$ denotes the sequential multiplication of the corresponding elements, $M_c(f^2)$ denotes the reinforcement of feature f^2 based on the channel attention mechanism, and $M_s(f^2)$ denotes the reinforcement of feature f^2 based on the spatial attention mechanism, as shown in Figure 4.

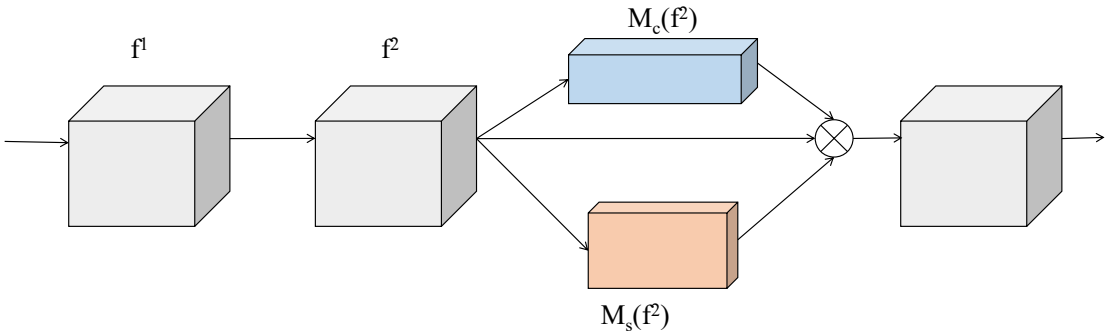

Figure 4. The attention mechanisms framework

Each channel of the feature can be treated as a feature detector, and compression of the input feature map in the spatial dimension can aggregate spatial information to determine the learnable information in the image. The input features are processed in the spatial dimension by both the maximum pooling layer and the average pooling layer, and then the processed features are fed into the shared multilayer perceptron, where the multilayer perceptron contains an implicit layer, and their outputs are summed to obtain the channel attention-enhancing feature $M_c(f^2)$ as shown in Equation (10):

$$M_c(f^2) = \sigma(MLP(f^c_{avg}) + MLP(f^c_{max})) \tag{10}$$

where $\sigma(\cdot)$ denotes the Sigmoid activation function, $MLP(\cdot)$ denotes the multilayer perceptron, f_{avg}^{c} is the feature obtained by applying the average pooling layer on the spatial dimension to feature f^2, and f_{max}^{c} is the feature obtained by applying the maximum pooling layer on the spatial dimension to feature f^2.

The difference with channel attention is that spatial attention focuses more on the location of the information, which can highlight the effective information region of the feature and get the local information of the image. The average pooling layer and maximum pooling layer operations are performed on the input features in the channel dimension respectively, and the output features are spliced together to generate the spatial attention reinforced feature $M_s(f^2)$ by a 7×7 convolutional layer as shown in Equation 11:

$$M_s(f^2) = \sigma(conv(f_{avg}^{s} \cup f_{max}^{s})) \tag{11}$$

where $\sigma(\cdot)$ denotes the Sigmoid activation function, $conv(\cdot)$ denotes the convolutional layer operation, $\cup$ denotes the feature splicing operation, f_{avg}^{s} is the feature obtained after applying the average pooling layer on the channel dimension to feature f^2, and f_{max}^{s} is the feature obtained after applying the maximum pooling layer on the channel dimension to feature f^2.

Based on spatial attention and channel attention, the input feature f^2 is processed to obtain the reinforced feature f^3 containing information that can stimulate emotion. Then the corresponding position of each channel in the reinforced feature f^3 is summed to obtain the feature f^4, which is the emotion mapping map. The original feature f^1 contains the global information of the picture, and the feature f^4 contains the local information of the picture, and the feature f^4 is coupled with the feature map on each channel in the original feature f^1 to obtain the feature map $U = [U_1, U_2, \ldots, U_n]$, where $U_i = f^4 \otimes f^1$. Then the original feature f^1 is spliced with the coupled feature $\bigcup$ to obtain feature f^5, which is used as an input to the sentiment classifier to predict the sentiment label.

The features f^4 are passed through the global average pooling layer and the fully connected layer to generate the feature vector d_{cls}. The probability distribution of the image sentiment is computed by the Softmax function, and the loss function for this task is shown in Equation (12):

$$L_{cls}(v, c; X) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} (y_i = c) \log(F^C(X_i|v)) \tag{12}$$

where N denotes the total number of images in the sentiment dataset, C denotes the number of categories of sentiment labels, y_i denotes the sentiment labels of X_i images, v denotes the learnable parameters of the model, and $F^C(X_i|v)$ denotes the probability that the sentiment of the i-th image in the dataset is predicted to be labeled as $c(s) = 1$ when the condition s is true, and 0 otherwise.

3.2. **Joint training processing.** The model in this paper handles two tasks simultaneously during training: predicting the rotation angle of an image and the sentiment labeling of an image. The loss function that unites these two tasks is shown in Equation (13). The network parameters are optimized using stochastic gradient descent to minimize the total loss function of the model.

$$L = \lambda L_{rot} + L_{cls} \tag{13}$$

where λ denotes the balance coefficient between the sentiment classification task and the self-supervised assistance task. This model increases the semantic diversity of features through the image rotation self-supervised assistive task, considers global and local

features based on the attention mechanism, and generates the sentiment mapping graph, which leads to enhanced discriminative features and improved model performance.

4. Experiment.

4.1. Experimental data set.

In this paper, we conducted experiments on emotion classification on the FI dataset, which contains 23,308 images from Flickr and Instagram platforms, annotated by 225 participants from AMT, using 8 emotion labels, namely fear, sadness, anger, disgust, satisfaction, excitement, entertainment, and awe.

4.2. Evaluation criteria.

For the sentiment categorization experiments, this paper uses accuracy as an evaluation criterion, defined as follows:

$$Accuracy = \frac{\sum_{i=1}^{c} N_i^c}{\sum_{i=1}^{c} N_i^s} \tag{14}$$

where c is the number of sentiment categories, N_i^c denotes the number of correctly predicted pictures in the i-th sentiment category, and N_i^s denotes the number of pictures in the i-th sentiment category.

4.3. Analysis of experimental results.

4.3.1. *Emotion classification experiment.* We first conduct a comparative experiment on the FI dataset for image sentiment classification, and the experimental results are shown in Table 1.

Table 1. Emotional accuracy comparison test (%)

ISLAMNet	PAEF	JCDL	SPN	WILDCAT	WSCNet
70.7	46.1	66.8	66.6	67.0	70.1

From Table 1, it can be found that the method in this paper outperforms the comparative benchmark models on the multi-categorization dataset. Compared with the WSCNet method, the accuracy of this chapter's method on the FI dataset is improved by about 0.6%, which proves that the model can extract more semantically informative features with better discriminative properties for the visual emotion recognition task.

The confusion matrix of ISLAMNet proposed in this paper on the FI dataset is shown in Figure 5. In the confusion matrix on the FI dataset, the emotion classification accuracies of fear and anger are the lowest, 0.39 and 0.47, respectively, and these two emotion pictures are often incorrectly categorized as other emotions in the negative emotions.

4.3.2. *Rotation angle experiment.* We first analyze the picture rotation angles in the self-supervised assisted task by conducting experiments on the FI dataset to analyze the effects of four different rotation tasks on the model accuracy: (a) rotating the pictures clockwise by 0° and 180°, respectively, (b) rotating the pictures clockwise by 90° and 270°, respectively, (c) rotating the pictures clockwise by 0°, 90°, 180°, and 270°, (d) rotating the pictures clockwise by 0°, 45°, 90°, 135°, 180°, 225°, 270°, and 315°.

Except for the changes in the rotation angle of the pictures, the rest of the parameter settings were the same as those in the experiment of sentiment categorization in Table 1. The experimental results are shown in Table 2.

We found that the classification accuracy of the rotation task (c) is the highest, so in the main experiment, we set the rotation angles of the pictures to 0°, 90°, 180°, 270°, followed by task (d), and the classification accuracy of task (b) is lower than that of task (a) for both tasks (a) and (b), even though both of them rotate the pictures twice.

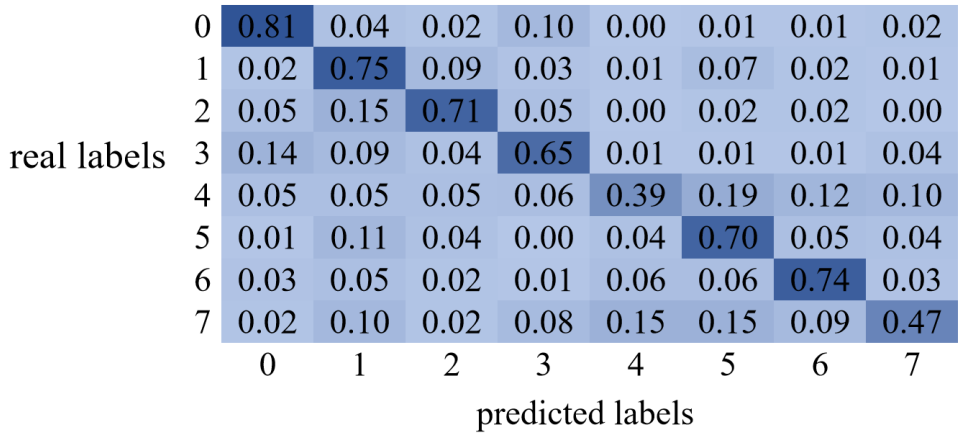

Figure 5. Confusion matrix

Table 2. Experimental comparison of different rotation angles

Task	Number of rotations	Rotation angle (°)	Classification Accuracy
a	2	0,180	57.1
b	2	90,270	55.1
c	4	0,90,180,270	59.4
d	8	0,45,90,135,180,225,270,315	58.1

After analyzing the results, we believe that the reason is: task (a) and (b) only rotate the image twice, and the model cannot learn enough useful semantic information by learning these two angles; and the image used for emotion recognition is the image rotated by 0°, so the self-supervised task brings less auxiliary supervised information to the backbone feature extraction network for 90° and 270° rotations, which makes the accuracy of task (a) and (b) the lowest. Task (b) has the lowest accuracy; and task (d), although the number of rotations is increased, the small gap between the rotation angles increases the classification difficulty of the rotation task, which makes the semantic feature extraction not accurate enough and leads to a decrease in the classification accuracy of the whole model.

4.3.3. *Affective region detection.* We analyze the performance of the method for emotion region detection. The FI dataset contains emotion salient map annotations and eye-tracking data; therefore, we analyze the performance of emotion mapping maps generated by the method on the FI dataset. Six evaluation metrics, including two Area Under Curve (AUC) variables and four similarity metrics, were used for the features as emotion mapping maps. Among them, the AUC-J variable and the AUC-B variable [25] use the emotion mapping map as a binary classifier to reduce the effect of the center bias, and the linear correlation coefficient [26], the histogram cross kernel [27], the moving distance [28], and the relative entropy are used to measure the similarity between the emotion mapping map and the eye gaze map. For AUC-J, AUC-B, CC, and SIM metrics, higher values indicate better performance, whereas for KL and EMD metrics, lower values indicate better performance.

In this paper, we compare with five baseline methods, and the experimental parameter settings are the same as those in the sentiment classification experiment on the multi-categorical dataset in Table 1. The experimental results are shown in Table 3.

Table 3. Experimental comparison of different rotation angles

Method	AUC-J	AUC-B	CC	SIM	KL	EMD
Objectness	0.61	0.56	0.17	0.31	7.51	5.04
GBVS	0.80	0.66	0.46	0.47	5.96	4.59
IttiKoch	0.73	0.63	0.37	0.43	2.09	3.16
WILDCAT	0.55	0.52	0.37	0.32	1.66	4.52
WSCNet	0.76	0.64	0.48	0.48	1.23	3.63
ISLAMNet	0.71	0.69	0.49	0.51	1.56	3.35

From Table 3, it can be seen that the SLAMNet model proposed in this paper outperforms the five other methods in AUC-B, CC, and SIM evaluation metrics and is close to the best result in KL and EMD evaluation metrics. This indicates that the overall performance of the sentiment mapping map extracted by the method in this paper is better.

Experiments show that the method in this paper effectively improves the semantic richness in the features. The self-supervised assisted task by picture rotation enables the backbone feature extraction network to recognize the semantic information in the pictures and generate the sentiment mapping map based on the attention mechanism to strengthen the extracted features. The model in this paper performs better and the sentiment classification accuracy is better than the comparative baseline model.

5. **Conclusion.** In this paper, we propose a visual emotion recognition method that integrates self-supervised learning and attention mechanism. Firstly, the image features are extracted by convolutional neural network, and the self-supervised auxiliary task of image rotation is added to enrich the semantic information of the features, and then the features are strengthened based on spatial attention and channel attention to improve the discriminative properties of the features. Experiments are conducted on the dataset, and the model's sentiment recognition accuracy is better than that of the comparative benchmark models, illustrating the effectiveness of the model. Compared to the sentiment recognition model without self-supervised learning, the model has higher sentiment classification accuracy, indicating that self-supervised learning can effectively utilize the close connection between sentiment and semantic information in pictures. The model in this paper performs well on the dataset and outperforms existing known methods. The research in this paper will contribute to the application of visual emotion recognition in the field of human-computer interaction.

REFERENCES

[1] B. Myers, J. Hollan, I. Cruz, S. Bryson, D. Bulterman, T. Catarci, W. Citrin, E. Glinert, J. Grudin, and Y. Ioannidis, "Strategic directions in human-computer interaction," *ACM Computing Surveys (CSUR)*, vol. 28, no. 4, pp. 794-809, 1996.

[2] J. O. Wobbrock, and J. A. Kientz, "Research contributions in human-computer interaction," *Interactions*, vol. 23, no. 3, pp. 38-44, 2016.

[3] J. Kammersgaard, "Four different perspectives on human–computer interaction," *International Journal of Man-Machine Studies*, vol. 28, no. 4, pp. 343-362, 1988.

[4] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging recent advances in deep learning for audio-visual emotion recognition," *Pattern Recognition Letters*, vol. 146, pp. 1-7, 2021.

[5] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 60-75, 2017.

[6] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, no. 2, 401, 2018.

[7] G. J. Lewis, C. E. Lefevre, and A. W. Young, "Functional architecture of visual emotion recognition ability: A latent variable approach," *Journal of Experimental Psychology: General*, vol. 145, no. 5, 589, 2016.
[8] T. Rao, X. Li, and M. Xu, "Learning multi-level deep representations for image emotion classification," *Neural Processing Letters*, vol. 51, pp. 2043-2061, 2020.
[9] K. Song, T. Yao, Q. Ling, and T. Mei, "Boosting image sentiment analysis with visual attention," *Neurocomputing*, vol. 312, pp. 218-228, 2018.
[10] D. She, J. Yang, M.-M. Cheng, Y.-K. Lai, P. L. Rosin, and L. Wang, "WSCNet: Weakly supervised coupled networks for visual sentiment classification and detection," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1358-1371, 2019.
[11] J. Yang, D. She, M. Sun, M.-M. Cheng, P. L. Rosin, and L. Wang, "Visual sentiment prediction based on automatic discovery of affective regions," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2513-2525, 2018.
[12] Z. Li, Y. Fan, W. Liu, and F. Wang, "Image sentiment prediction based on textual descriptions with adjective noun pairs," *Multimedia Tools and Applications*, vol. 77, pp. 1115-1132, 2018.
[13] T. Yamamoto, S. Takeuchi, and A. Nakazawa, "Image emotion recognition using visual and semantic features reflecting emotional and similar objects," *IEICE Transactions on Information and Systems*, vol. 104, no. 10, pp. 1691-1701, 2021.
[14] J. Zhang, X. Liu, Z. Wang, and H. Yang, "Graph-based object semantic refinement for visual emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 3036-3049, 2021.
[15] T.-Y. Wu, H. Li, S. Kumari, and C.-M. Chen, "A Spectral Convolutional Neural Network Model Based on Adaptive Fick's Law for Hyperspectral Image Classification," *Computers, Materials & Continua*, vol. 79, no. 1, pp. 19-46, 2024.
[16] F. Zhang, T.-Y. Wu, J.-S. Pan, G. Ding, and Z. Li, "Human motion recognition based on SVM in VR art media interaction environment," *Human-centric Computing and Information Sciences*, vol. 9, 40, 2019.
[17] F. Zhang, T.-Y. Wu, and G. Zheng, "Video salient region detection model based on wavelet transform and feature comparison," *EURASIP Journal on Image and Video Processing*, vol. 2019, 58, 2019.
[18] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, no. 4, pp. 193-202, 1980.
[19] M. C. Schiappa, Y. S. Rawat, and M. Shah, "Self-supervised learning for videos: A survey," *ACM Computing Surveys*, vol. 55, pp. 1-37, 2023.
[20] R. Krishnan, P. Rajpurkar, and E. J. Topol, "Self-supervised learning in medicine and healthcare," *Nature Biomedical Engineering*, vol. 6, no. 12, pp. 1346-1352, 2022.
[21] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 857-876, 2021.
[22] Y. Xie, Z. Xu, J. Zhang, Z. Wang, and S. Ji, "Self-supervised learning of graph neural networks: A unified review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2412-2429, 2022.
[23] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533-536, 1986.
[24] A. Saeed, T. Ozcelebi, and J. Lukkien, "Multi-task self-supervised learning for human activity detection," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 2, pp. 1-30, 2019.
[25] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 55-69, 2012.
[26] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision Research*, vol. 47, no. 19, pp. 2483-2498, 2007.
[27] M. J. Swain, and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11-32, 1991.
[28] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, pp. 99-121, 2000.