Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.036975
Article

Image Emotion Classification Network Based on Multilayer Attentional Interaction, Adaptive Feature Aggregation

Xiaorui Zhang[1,2,3,*], Chunlin Yuan[1], Wei Sun[3,4] and Sunil Kumar Jha[5]

[1]Engineering Research Center of Digital Forensics, Ministry of Education, Jiangsu Engineering Center of Network Monitoring, School of Computer and Software, Nanjing University of Information Science & Technology, Nanjing, 210044, China
[2]Wuxi Research Institute, Nanjing University of Information Science & Technology, Wuxi, 214100, China
[3]Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science & Technology, Nanjing, 210044, China
[4]School of Automation, Nanjing University of Information Science & Technology, Nanjing, 210044, China
[5]Adani University, Ahmedabad, Gujarat, India
*Corresponding Author: Xiaorui Zhang. Email: zxr365@126.com
Received: 18 October 2022; Accepted: 30 January 2023

Abstract: The image emotion classification task aims to use the model to automatically predict the emotional response of people when they see the image. Studies have shown that certain local regions are more likely to inspire an emotional response than the whole image. However, existing methods perform poorly in predicting the details of emotional regions and are prone to overfitting during training due to the small size of the dataset. Therefore, this study proposes an image emotion classification network based on multilayer attentional interaction and adaptive feature aggregation. To perform more accurate emotional region prediction, this study designs a multilayer attentional interaction module. The module calculates spatial attention maps for higher-layer semantic features and fusion features through a multilayer shuffle attention module. Through layer-by-layer up-sampling and gating operations, the higher-layer features guide the lower-layer features to learn, eventually achieving sentiment region prediction at the optimal scale. To complement the important information lost by layer-by-layer fusion, this study not only adds an intra-layer fusion to the multilayer attention interaction module but also designs an adaptive feature aggregation module. The module uses global average pooling to compress spatial information and connect channel information from all layers. Then, the module adaptively generates a set of aggregated weights through two fully connected layers to augment the original features of each layer. Eventually, the semantics and details of the different layers are aggregated through gating operations and residual connectivity to complement the lost information. To reduce overfitting on small datasets, the network is pre-trained on the FI dataset, and further weight fine-tuning is performed on the small dataset. The experimental results on the FI, Twitter I

and Emotion ROI (Region of Interest) datasets show that the proposed network exceeds existing image emotion classification methods, with accuracies of 90.27%, 84.66% and 84.96%.

Keywords: Attention mechanism; emotional region prediction; image emotion classification; transfer learning

1 Introduction

Psychological studies show that people are easily stimulated by visual content, especially images, to produce corresponding emotions [1]. With the accelerated pace of life and the development of the Internet, more and more people like to express their emotional states and opinions by sharing images on social media platforms such as Weibo and WeChat. If the emotional information contained in images can be fully applied to human emotion classification, this will contribute to business behavior and psychological studies such as accurate advertising placement, content delivery, and game scene modeling. Therefore, in recent years, many domestic and international scholars have conducted a lot of studies in the context of image emotion classification [2–5].

Before deep learning became popular, scholars mainly studied hand-crafted features such as color and texture for emotion classification. However, handcrafted features [6,7] are labor-intensive and time-consuming. In addition, such lower-layer features fail to adequately express advanced attributes such as emotion. In contrast, neural networks can automatically learn deeper features and can characterize advanced attributes such as emotion more effectively. At the same time, many existing studies have shown that emotional changes in people are often captured from certain regions of the image [8,9], which are called emotional regions. Therefore, different methods have been proposed to improve classification accuracy by investigating emotional regions [10,11]. However, they still have certain limitations: (i) People's attention is influenced by features at different layers, most current methods rely on higher-layer semantic features for emotional region prediction, but neglect lower-layer fine-grained features. (ii) The layer-by-layer fusion manner doesn't exchange information directly at the layers that are far away, so the important information captured at the higher-layer may be lost in the entire gradual process [12]. (iii) To ensure high accuracy, sufficient data is needed to train the model. However, due to social privacy and annotation issues, several commonly used public image emotion datasets are small, making the models prone to overfitting when trained.

Therefore, this study proposes an image emotion classification network based on multilayer attentional interaction and adaptive feature aggregation. To make full use of the extracted multilayer features for more accurate emotional region prediction, this study designs a multilayer attentional interaction module, which is inspired by the work of Yang et al. [13] and Wu et al. [14]. The module acts between multilayer outputs, which pass channel and spatial attention information from top to bottom and enable higher-layer features to guide lower-layer features learning. Emotional region prediction is achieved by progressively integrating multilayer attention at the optimal scale. To complement the information that may be lost in the multilayer attentional interaction module, this study proposes an adaptive feature aggregation module, which is inspired by the work of Sun et al. [15]. This module adaptively predicts a set of weights based on the importance of the features in each layer, augmenting the original features in each layer. The information is complemented by aggregating the semantics and details of different layers through residual connections with the emotional region features. To reduce overfitting on small datasets, the network is pre-trained on the FI dataset, and further weight

fine-tuning is performed on the small dataset. In addition, weight decay, data augmentation, and two dropout operations with a probability of 0.5 are used to jointly reduce overfitting.

In summary, our contributions are mainly as follows:

(i) This study designs a multilayer attentional interaction module. The module computes a spatial attention graph from higher-layer semantic features and uses intra-layer fusion to enhance key information. Then, layer-by-layer up-sampling and gating operations are used to enable higher-layer features to guide lower-layer features to learn, ultimately achieving the emotional region prediction at the optimal scale.

(ii) This study designs an adaptive feature aggregation module. The module uses global average pooling to connect channel information of all layers and adaptively generates a set of aggregated weights through two fully connected layers, enhancing the original features of each layer. Eventually, the semantics and details of different layers are directly aggregated through gating operation and residual connections to complement the information lost in the multilayer attentional interaction module.

(iii) In this study, the network is pre-trained on the FI dataset and further weight fine-tuning is performed on the small dataset. A combination of weight decay, data augmentation, and two dropout operations are also incorporated to reduce overfitting, thereby ensuring high accuracy.

The rest of the article is arranged as follows. Section 2 describes the related work. Section 3 describes the proposed model. Section 4 describes the experimental environment and the dataset, gives the experimental protocol, and analyses the results. Section 5 discusses the advantages and disadvantages of the proposed model and gives directions for future study.

2 Related Work

This section introduces the works related to our model, such as emotional region prediction, attention mechanisms, and transfer learning.

2.1 Emotional Region Prediction

What inspires emotion in the image? Many studies in recent years have shown that only certain regions of an image tend to inspire emotion, while other unimportant regions should be ignored. These remind people that when predicting image emotion, we can not only use global information but also add important inferential information of emotion region to assist image emotion classification [16]. How to accurately capture the emotional regions of an image has become an urgent problem.

Table 1: Pros. and cons. of image emotion classification models

Author	Particular year	Pros.	Cons.
Yang et al. [13]	2018	Classify emotions according to local image features	Simple weighting through cross-space pool module
Wu et al. [14]	2019	Use the object detection module to determine local areas	Ignore the contribution of different layers
Yao et al. [17]	2019	Conduct polarity and emotion specific attention on the lower layers and higher layers	Lack of guided learning between multilayer features in the emotional region prediction process

(Continued)

Table 1: Continued

Author	Particular year	Pros.	Cons.
Rao et al. [10]	2020	Obtain local features through Feature Pyramid Network (FPN) and ResNet, and connect local features with global features	Use serial networks with high computational complexity; Important information may be lost during fusion
Priya et al. [18]	2020	Combine the extracted high-layer features and low-layer features	Fusion accords to the same weight, ignoring the human attention mechanism
Qu et al. [19]	2021	Based on the alterable scale and multi-level local regional emotional affinity analysis under the global perspective	Although FPN is improved, it is still a parallel structure; Important information may be lost during fusion

In recent years, emotion-region-based image emotion classification methods have achieved encouraging performance improvements on many image emotion datasets. Table 1 shows the pros. and cons. of image emotion classification models. Peng et al. [20] proposed that emotions are often induced by specific regions and presented the Emotion ROI dataset. Yang et al. [13] proposed WSCNet, which weights the final output features by a cross-space pooling module and classifies emotions according to local image features. Wu et al. [14] suggested a target detection module to determine whether to use local regions. However, these methods all only use the highest-layer feature to identify the emotional regions and ignore the contribution of different layers of features to image emotion representation, thereby, limiting their performance.

To further improve the model performance, Rao et al. [10] used FPN to fuse the single layer of features from the ResNet network via a layer-by-layer manner to obtain multilayer features. Then, the emotional region information in the multilayer features is extracted by Faster Region-based Convolutional Neural Network (Faster R-CNN) and the obtained information is used as local features to connect with the global features. Yao et al. [17] conducted polarity and emotion-specific attention on the lower layers and higher layers, respectively. Priya et al. [18] combines the extracted high-level features and low-level features with equal weight. Qu et al. [19] proposed a multi-level context pyramid network (MCPNet) for visual sentiment analysis by combining local and global representations. These methods make full use of different layers of features to represent image emotion. However, they still have the following limitations: (1) these methods used two serial networks to achieve multilayer feature extraction, which is labor-intensive and time-consuming. Additionally, the obtained multilayer features have a low resolution; (2) the Faster R-CNN extracts emotional regions only separately using the multilayer features from FPN and lacks guided learning between multilayer features in the emotional region prediction process; (3) the layer-by-layer feature fusing will lead to indirect information propagation between distant layers, making useful information captured at the higher layer to be lost in a progressive process.

Therefore, this study proposes to use a parallel High Resolution network (HRNet) [21] as the backbone network to directly extract multilayer original features, which can accelerate network training and ensure that all layers of the network have relatively high-resolution features. Adding a multilayer attentional interaction module after the backbone network and combining layer-by-layer fusion with emotional region prediction enables higher-layer features to guide lower-layer features learning. Therefore, the emotional region prediction features with optimal spatial resolution can be

obtained at the lowest-layer, making the emotional regions more refined. Further details can be found in Section 4.2. In addition, an adaptive feature aggregation module is designed. This module adaptively predicts a set of weights based on the importance of features in each layer, enhancing the original features in each layer. By aggregating the semantics and details of different layers, the lost information during the layer-by-layer fusion process is complemented. Further details can be found in Section 4.3.

2.2 Attention Mechanism

The attention mechanism enables the network to focus on a specific region and enhance the features of these regions, which meets the need of the emotional region prediction task. Currently, there are three types of attention mechanisms used in computer vision, namely channel attention, spatial attention, and hybrid attention.

Channel attention focuses on modeling the correlations between channels and assigning weights to each channel according to its importance. Hu et al. [22] proposed the Squeeze and Excitation Networks (SE Net). SE Net improves network accuracy by modeling the correlations between feature channels and weighting the important ones. SE Net reduces the high complexity caused by capturing all inter-channel correlations by dimensionality reduction, but dimensionality reduction interferes with the learning of channel attention. Wang et al. [23] found that capturing all inter-channel correlations is inefficient and unnecessary. Therefore, they proposed an Efficient Channel Attention (ECA) with local cross-channel interaction, which captures local cross-channel interaction information by considering each channel and its k neighbors. This method ensures modeling efficiency and computational effectiveness.

In contrast, the spatial attention mechanism extracts the spatial attention matrix and assigns more weight to the more important spatial pixels to identify the regions that need to be focused on.

Hybrid attention is a weighted fine-tuning of the feature map at both the channel level and spatial level. The commonly used hybrid attention CBAM [24] is effective, but it has a large amount of calculation. To reduce model complexity and calculation, Zhang et al. [25] proposed a lightweight yet efficient Shuffle Attention (SA) module, which divides channel dimensions into sub-features. For each sub-feature, shuffle units were used to construct both channel attention and spatial attention. The attention module is designed with an attention mask at all positions, which suppresses possible noise and highlights the regions with the correct semantic features. In contrast, the SA module is more accurate than previous hybrid attention methods and contains fewer parameters.

In this study, the SA module is added after each layer of the output of the backbone network to focus on the important regions of different layer features, which obtains an enhanced pixel-by-pixel emotional region prediction at the optimal scale by progressively integrating multilayer attention. At the same time, this study proposes to add an ECA module without dimension reduction after the fused features to strengthen important channels again, so that the network learns more effective features.

2.3 Transfer Learning

Deep learning requires large amounts of training data to understand the underlying patterns of the data [26]. However, due to social privacy and labeling difficulties, several commonly used publicly available image emotional datasets are small. Therefore, many models are undertrained, which prevents these models from achieving optimal performance in image emotion classification. Previous studies [27–29] have shown that neural networks based on transfer learning can effectively solve the problems caused by insufficient data.

Wei et al. [27] proposed a transfer learning model combining convolutional and recurrent neural networks for emotion analysis. The model uses a convolutional neural network to acquire features and a recurrent neural network to learn the feature representation. Then the model is trained in the source domain and parameters are tuned through iterative calculation. When applying the parameters to the target domain, the model is further fine-tuned based on the target domain data to achieve better classification results. Zhang et al. [28] exploited the weight parameters trained on the ImageNet database to initialize the MobileNetV2 network, and then retrained the model based on the CT image data set provided by Kaggle. Dai et al. [29] took the parameters that are trained from unlabeled data as the initial parameters for next-stage model training. The pre-training method effectively combined the advantages of unsupervised and supervised learning and achieved an average classification accuracy of 92.80% for text emotion classification. Currently, most models for emotion classification are trained and tested on a single dataset. By means of transfer learning, these above models are adequately trained, and the test results are significantly better than those without using transfer learning, which can solve the problem of insufficient data.

Inspired by the above studies, this study proposes to use the FI dataset to pre-train the network, and further fine-tune the network on two small datasets, Twitter I and Emotion ROI. This will solve the overfitting caused by insufficient data.

3 Proposed Method

In this section, this study proposes an image emotion classification network based on multi-layer attentional interaction and adaptive feature aggregation. To predict emotional regions more accurately, this study designs a multilayer attentional interaction module after a multilayer feature extraction module. In addition, to compensate the multilayer attentional interaction module for the lost information, an adaptive feature aggregation module is designed, and the two modules are merged by residual connections. The following will describe the details of the proposed method.

3.1 Structure

The proposed network structure is shown in Fig. 1. This study uses a HRNet with four layers as the backbone network, which continuously exchanges information between multilayer features to maintain high resolution. To make full use of the multilayer features extracted by the backbone network for more accurate emotional region prediction, this study designs a multilayer attentional interaction module. The highest-layer features are firstly computed by Shuffle Attention to obtain the highest-layer spatial attention features, followed by up-sampling and gating operations to obtain the second-highest-layer gated fusion features. Repeat the above interaction process for the gated fusion features obtained at each layer, until the gated fusion features with the optimal spatial resolution are obtained at the bottom layer. To reduce the information loss due to layer-by-layer fusion, this study gates the obtained spatial attention features of each layer with the original features of that layer. To make the features more effective, the lowest-layer gated fusion features are fed into the undescended ECA module to obtain the emotional region features.

Important information is probably lost during the layer-by-layer fusion process. To compensate the multilayer attentional interaction module for the lost information, this study proposes an adaptive feature aggregation module. The module adaptively predicts a set of weights based on the importance of the features in each layer, augmenting the original features in each layer and aggregating the semantics and details of different layers. A 1×1 convolutional layer is added after each layer of HRNet for dimensionality reduction. Then, global average pooling is used at each layer to compress

spatial information to further connect channel information across all layers. At the same time, a simple gating operation is performed on the aggregated features, to make the module learns a fused weight. Using the fused weight, this study dynamically enhance the original features in each layer. The adaptively enhanced multilayer features are used as the global features and are dotted with the emotional region features. Then, make a residual connection between the dotted product and global features. The connected features are passed through the global average pool to obtain the emotion classification features, and the obtained emotion classification features are fed into the fully connected layer. Finally, the emotion is classified using a SoftMax classifier.

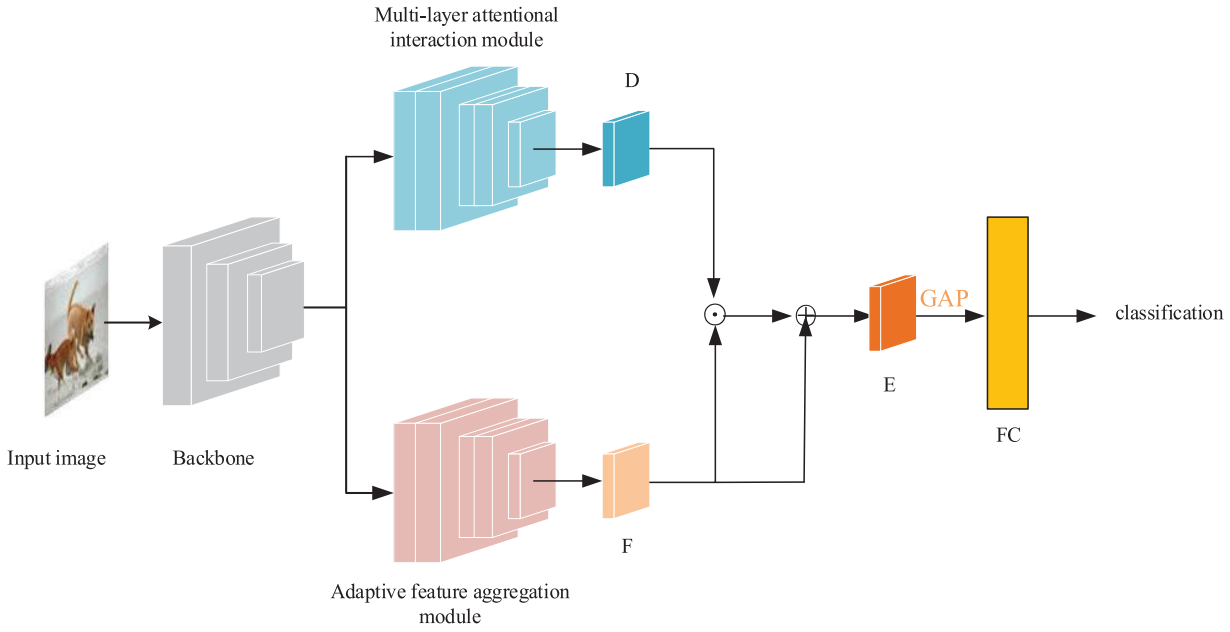

Figure 1: Image emotion classification network based on multilayer attentional interaction and adaptive feature aggregation. In the figure, D denotes emotional region features, F denotes global features, and E denotes integrated emotion features

3.2 Multilayer Attention Interaction Module

Unlike the traditional method such as a simple connection of multilayer input features [30], this study designs a multilayer attention interactional module to make full use of the multilayer features extracted by HRNet. The module enables the higher-layer features to guide the lower-layer features for predicting emotional regions at the bottom layer with the highest resolution. The structure of the module is shown in Fig. 2 and its implementation is detailed described below.

This study uses HRNetV2-w48 as the backbone network to extract multilayer features. In this study, C_i denotes the i-th layer original features from top to bottom, $i = 1, 2, 3, 4$; A_j denotes the spatial and channel attention features of layer j, $j = 1, 2, 3$; O_k denotes the gated fusion features of layer k, $k = 2, 3, 4$. Firstly, the highest-layer features C_1 calculate the corresponding spatial and channel attention features A_1 through SA. Multiply A_1 with C_1, and up-sample the result to the second-highest-layer resolution. Then multiply the up-sampled features with C_2 by the corresponding element to obtain the second-highest-layer gated fusion features O_2. Next, the spatial attention features A_2 of the second-highest-layer gated fusion features O_2 are computed by SA. A_2 is multiplied with O_2 by the

corresponding element, and the result is up-sampled and gated with C_3 for fusion. The above process is repeated until the lowest-layer gated fusion features O_4 are obtained. This multilayer attentional interaction operation can be described by the following equation:

$$A_1 = \sigma\left(W^{7\times7}\left([AvgPool\left(C_1\right); MaxPool\left(C_1\right)]\right)\right) \tag{1}$$

$$O_2 = up(A_1 \otimes C_1) \otimes C_2 \tag{2}$$

$$A_m = \sigma\left(W^{7\times7}\left([AvgPool\left(O_m\right); MaxPool\left(O_m\right)]\right)\right) \tag{3}$$

$$O_{m+1} = up(A_m \otimes C_m) \otimes C_{m+1} \tag{4}$$

$$D = \sigma\left(Conv_{1\times1}(ECA(O_4))\right) \tag{5}$$

where $\sigma()$ denotes the sigmoid activation function, $W^{7\times7}$ denotes the weight of the 7×7convolution, $avgPool()$ denotes the average pooling operation, $MaxPool()$ denotes the maximum pooling operation, $up()$ denotes bilinear interpolation up-sampling, $\otimes$ denotes the gating operation (Corresponding elements are multiplied together), and m denotes the layer m, $Conv_{1\times1}()$ denotes 1×1 convolution, ECA() denotes efficient channel attention module.

The spatial attention features A_1 from the highest-layer features C_1 are calculated by Eq. (1). The calculation process is as follows. This uses global maximum pooling and global average pooling to compress in the channel domain, and obtain the global maximum pooling features and the global average pooling features from C_1. These two pooled features are connected along the channels to obtain the features with two channels. Then a 7×7 convolution is used to fuse the two pooled features, which can reduce the channel number to 1. Finally, the spatial attention and channel features A_1 are output through the sigmoid activation function. After implementing the above four equations, the final layer of the gated fusion features O_4 contain both higher-layer semantic features and lower-layer high resolution detail features. To further strengthen the channel features of the feature map to improve the accuracy, the lowest-layer gated fusion features are fed into an ECA module. Then, the number of channels of the features is reduced to 1 by 1×1 convolution. Finally, a sigmoid activation function is used to normalize the emotional region features D.

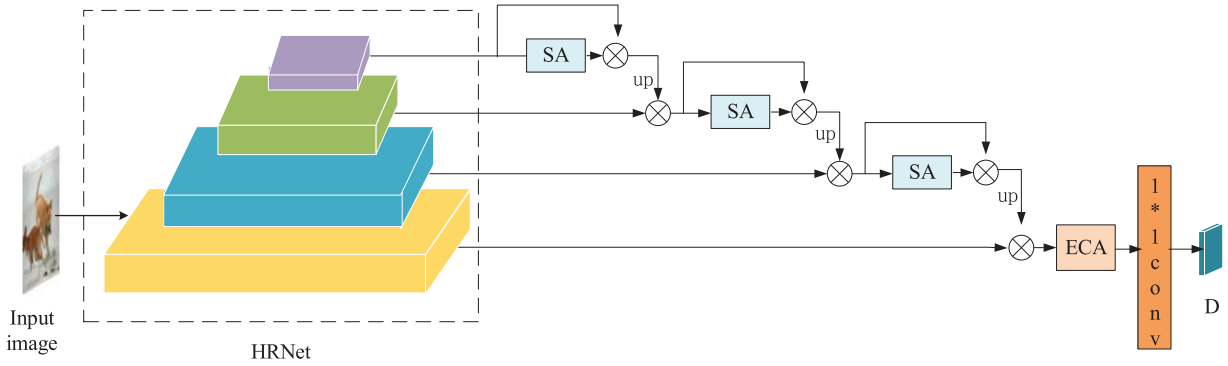

Figure 2: Multilayer attentional interaction module. In the figure, SA denotes Shuffle Attention, ECA denotes Efficient Channel Attention, D denotes emotional region features

3.3 Adaptive Feature Aggregation Module

Important information may lose during the layer-by-layer fusion process. To complement the lost information, this study proposes an adaptive feature aggregation module. The module allows information to be exchanged between multiple layers and adaptively generates a set of aggregated weights to enhance the original features of each layer. The structure of the module is shown in Fig. 3, and the implementation is described in detail below.

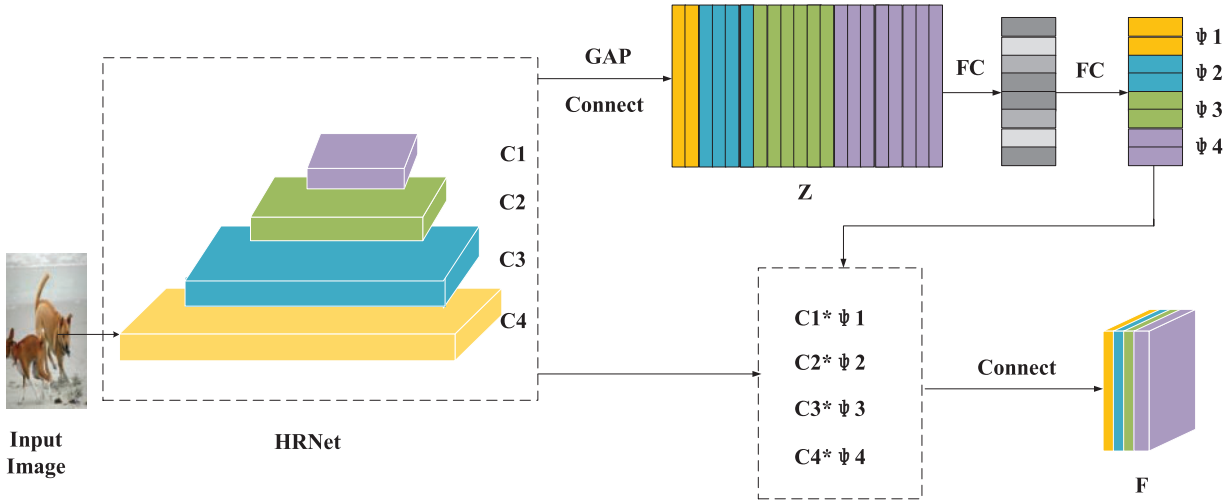

Figure 3: Adaptive feature aggregation module. In the figure, C_i denotes HRNet's features, Z denotes channel-level global features, F denotes global features

For each layer of HRNet's features C_i, $i = 1, 2, 3, 4$, this study adds a 1×1 convolutional layer to each layer for dimensionality reduction. Then, global average pooling is used at each layer to compress spatial information and connect channel information from all layers to obtain channel-level global features Z. To better aggregate the semantics and details of different layers, the module learns a hierarchy of adaptive aggregated weights $\psi \in R^{1\times4}$ via Eq. (6),

$$\psi = W_2\left(ReLU\left(W_1\left(Z\right)\right)\right) \tag{6}$$

where W_1 and W_2 are the weights of the two fully connected layers, and $ReLU()$ is the ReLU activation function.

Using the aggregated weights ψ, this model dynamically obtains the enhanced features $\widetilde{C}_i$ for each layer according to Eq. (7),

$$\widetilde{C}_i = C_i * \psi_i \tag{7}$$

where ψ_i is the i-th element in ψ, and $*$ denotes the scalar multiplication between C_i and ψ_i.

Subsequently, $\widetilde{C_{2\sim4}}$ is up-sampled by bilinear interpolation to the same resolution as $\widetilde{C}_1$, and then they are connected to generate the global features F. Formally, this process can be expressed as Eq. (8),

$$F = \widetilde{C}_1 \oplus up\left(\widetilde{C}_2\right) \oplus up\left(\widetilde{C}_3\right) \oplus up\left(\widetilde{C}_4\right) \tag{8}$$

where $\oplus$ is a series operation and $up()$ denotes bilinear interpolation up-sampling.

3.4 Classifier

To better combine the emotional region features with the global features, this study firstly dots D with F and then connects the result with the global features F to get the integrated emotion features E. The fusion process can be expressed by Eq. (9). Finally, the global average pooling output the emotion vector with dimension 2048. Then the emotion vector is fed into the fully connected layer (FC) to classify the emotion by the SoftMax classifier. In this image emotion classification model, the classifier uses three fully connected layers. The first and the second fully connected layer reduce the dimension and use dropout to prevent overfitting. The third fully connected layer reduces the dimension to the number of categories and doesn't use dropout operation, and outputs the emotion classification results.

$$E = D \otimes F + F \tag{9}$$

In this study, cross-entropy is used as the objective function, and the L2 regular term is added to further avoid overfitting the model. The network model is optimized by minimizing the objective function and the parameter is optimized by the gradient descent algorithm. The objective function is as follows:

$$L = -\frac{1}{N}\sum_{a=1}^{N}\left[\hat{y}_a log y_a + \left(1-\hat{y}_a\right) log\left(1-y_a\right)\right] + \lambda \parallel \theta \parallel^2 \tag{10}$$

where L is the objective function, a is the sample subscript, N is the number of samples, $\hat{y}_a$ is the label of the sample (positive class is 1 and negative class is 0), y_a is the probability of a positive prediction, and $\lambda \parallel \theta \parallel^2$ is the regular term.

3.5 Method Flow

The flow chart of the proposed method is shown in Fig. 4, including the following steps.

Step 1: Use random clipping and random horizontal flipping to improve the quality of the images.

Step 2: Build HRNetV2-w48 networks as the backbone of the model.

Step 3: Build a multilayer attentional interaction module that consists of a multilayer shuffle attention module, an intra layer and inter layer gating fusion module and an efficient channel attention network.

Step 4: Build an adaptive feature aggregation module to adaptively generates a set of aggregated weights through two fully connected layers to augment the original features of each layer.

Step 5: Use a Stochastic Gradient Descent optimizer and transfer learning to train network.

For the convenience of reading, abbreviations and symbols used in this study are shown in Tables 2 and 3 respectively.

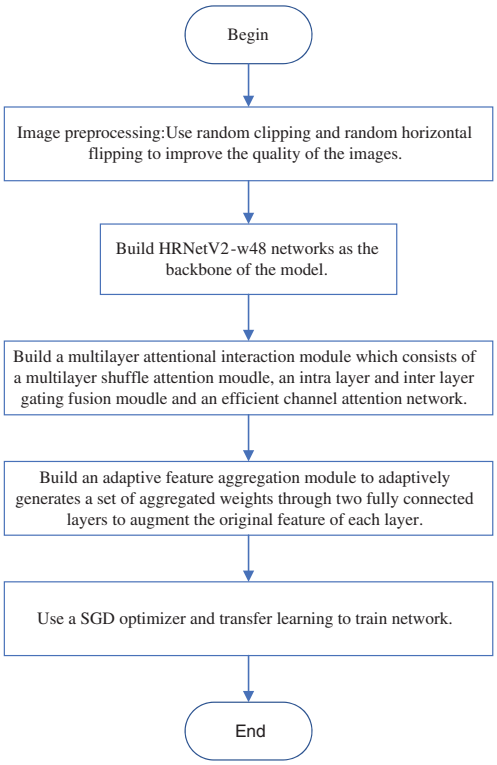

Figure 4: Flow chart of the proposed method

Table 2: Abbreviation

Full name	Abbreviation
Efficient Channel Attention	ECA
Feature Pyramid Network	FPN
Faster Region-based Convolutional Neural Network	Faster R-CNN
Fully Connected layer	FC
High Resolution network	HRNet
Region of Interest	ROI
Squeeze and Excitation Networks	SE Net
Shuffle Attention	SA

Table 3: Symbols

Symbol	Meaning
C_i	i-th original layer features, $i = 1, 2, 3, 4$
A_j	Spatial and channel attention features of layer j, $j = 1, 2, 3$

(Continued)

Table 3: Continued

Symbol	Meaning
O_k	Gated fusion features of layer k, $k = 2, 3, 4$
σ ()	Sigmoid activation function
$W^{7\times7}$	The weight of the 7×7convolution
avgPool ()	Average pooling
MaxPool ()	Maximum pooling
up ()	Bilinear interpolation up-sampling
$\otimes$	Gating operation
m	m-th layer
$Conv_{1\times1}$()	1×1 convolution
D	Emotional region features
Z	Channel-level global features
ψ	Adaptive aggregated weights
W_1	Weight of the first FC
W_2	Weight of the second FC
ReLU ()	ReLU activation function
$\widetilde{C}_i$	i-th layer enhanced features
$\oplus$	Series
F	Global features
E	Integrated emotion features
L	Objective function
a	Sample subscript
N	The number of samples
$\hat{y}_a$	Label
y_a	The probability of a positive prediction

4 Experiments

4.1 Experimental Setup

4.1.1 Equipment and Environment

The experiment and environment in this study are shown in Table 4.

Table 4: Experimental equipment and environment

Equipment	A computer with an Intel Core i7-10870 CPU A remote server consisting of 2 NVIDIA GeForce RTX 3090 GPU
Environment	pytorch 1.11 for model training and validation

4.1.2 Datasets

The number of images in the commonly used datasets, except for the FI dataset, is far from the requirement for training a multi-classification network. Therefore, three widely used image emotion datasets are chosen for the image emotion two-classification task in this study, namely the FI dataset [31], the Twitter I dataset [32] and the Emotion ROI dataset [20].

The FI dataset contains 21194 images, divided into eight emotion categories: pleasure, satisfaction, excitement, surprise, anger, disgust, fear and sadness. Although the FI dataset divides emotion into eight categories, it can still be practically summarized in two categories: positive and negative. This study follows the division method of previous studies [33], combining pleasure, satisfaction, excitement and surprise into a positive category with 15036 images, and anger, disgust, fear and sadness into a negative category with 6158 images. By doing so, the eight-classification task is transformed into a two-classification task. Like other studies [10,13] on emotion classification, the FI dataset is randomly divided into an 80% training set, a 5% validation set and a 15% test set in this experiment. Fig. 5 shows a selection of example images from the reclassified FI dataset.

Figure 5: A selection of sample images from the reclassified FI dataset

The Twitter I dataset contains 1269 images, divided into positive and negative categories. Since this dataset is small, this study randomly divides the dataset into an 80% training set and a 20% test set. Five-fold cross validation is used in the experiment and the final results were averaged over the five-fold cross validation.

The Emotion ROI dataset contains 1980 images, divided into six emotion categories, such as anger, disgust, fear, sadness, joy and surprise, with 330 images in each category. This study also summarizes the dataset into positive and negative categories, where the four categories of anger, disgust, fear and sadness are combined into the negative category, and joy and surprise are combined into the positive category. Five-fold cross validation is used in the experiment and the final results were averaged over the five-fold cross validation.

Table 5 shows the number of positive and negative images for the three data sets used in this study.

Table 5: The number of positive and negative images for FI, Twitter I and Emotion ROI

Datasets	Positive	Negative
FI	15036	6158
Twitter I	769	500
Emotion ROI	660	1320

4.1.3 Parameter Settings

Before training, the parameters of the HRNet are initialized to pre-trained parameters on the ImageNet classification task to improve accuracy and reduce training time.

Batchsize is set to 32. The network is trained by a Stochastic Gradient Descent optimizer. Momentum is set to 0.9 and weight decay is set to 0.001. For the FI dataset, the initial learning rate of the optimizer is set to 0.001. Note that every 30 epochs, the learning rate is reduced to one-tenth of the previous rate. This study trains 90 epochs on the FI dataset. For the two small datasets, Twitter I and Emotion ROI, to prevent overfitting, this study first pre-traines the network using the FI dataset and then fine-tunes it on these small datasets. At this time, the initial learning rate of the optimizer is set to 0.0001 and the other settings remain the same as for the FI dataset.

To reduce the risk of overfitting, in addition to setting the weight decay, data augmentation is used on the training data, which includes random cropping and random horizontal flipping. In the fully connected layer, two dropout operations with a probability of 0.5 are used to reduce overfitting. For greater clarity, the parameter settings in this article is shown in Table 6.

Table 6: The parameter settings

Parameter name	Parameter value
Batchsize	32
Momentum	0.9
Weight decay	0.001
Epochs	90
Initial learning rate (FI)	0.001 (every 30 epochs, the learning rate is reduced to one-tenth of the previous rate)
Initial learning rate (Twitter I and Emotion ROI)	0.0001 (every 30 epochs, the learning rate is reduced to one-tenth of the previous rate)
Dropout rate	0.5

4.1.4 Evaluating Indicator

For classification tasks, accuracy is the most common indicator to evaluate model performance. But the current distribution of positive and negative samples in most emotion datasets is unbalanced, so this study chooses accuracy and recall to comprehensively evaluate the model. Accuracy reflects the probability of the model classifying the emotion correctly in samples, and recall reflects the model's ability to correctly predict the positive samples. The specific calculation equations are as follows.

$$accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{11}$$

$$recall = \frac{T_P}{T_P + F_N} \tag{12}$$

T_P (true positive) denotes a positive sample predicted positive by the model. T_N (true negative) denotes a negative sample predicted negative by the model. F_N (false negative) denotes a positive sample predicted negative by the model. F_T (false positive) denotes a negative sample predicted positive by the model.

4.2 Experimental Results

4.2.1 Ablation Experiment

In this study, we conduct ablation experiments to analyze the importance of each module in the model. Table 7 lists the test results of the model's accuracy on each dataset when the corresponding module is removed. The accuracies are all declined to varying degrees, which demonstrates the effectiveness of our proposed modules.

Table 7: Ablation experimental results

Datasets/Ablation module	Normal	Without-multilayer attentional interaction	Without-adaptive feature aggregation
FI	90.27	88.36	89.57
Twitter I	84.66	81.49	83.55
Emotion ROI	84.96	81.57	83.74

Firstly, this study removes the multilayer attention interaction module. Only the global features obtained by the adaptive feature aggregation module are used for emotional classification, which doesn't consider the emotional regions of the image. The results show that the accuracies of the model degrades by 1.91, 3.17 and 3.39 on the FI, Twitter I and Emotion ROI datasets. This can demonstrate that the high-resolution emotional region features obtained by the multilayer attention interaction module can be effective in emotion classification. Further analysis shows that multi-level spatial channel attention fusion can bring more complex emotional region information, which is conducive to the following emotional classification.

Secondly, this study removes the adaptive feature aggregation module. Emotion classification is performed directly using the high resolution emotional region features obtained from the multilayer attentional interaction module, which doesn't complement the lost information. The experimental results show that the performance of the model degraded by 0.7, 1.11 and 1.22 on the FI, Twitter I and Emotion ROI datasets. The results show that the information lost in the layer-by-layer fusion process really affects the emotion classification. This problem can be effectively solved by aggregating multilayer features with adaptive weights.

Comparing the ablation results of two modules, the multilayer attentional interaction module is more effective in improving accuracy than the adaptive feature aggregation module. Further analysis shows that using multi-layer attention modules to enhance discriminative information is more effective in improving accuracy for image classification tasks.

In addition, to demonstrate the superiority of transfer learning, experiment is conducted on the model without transfer learning. The results are shown in Table 8.

The experimental results show that the model based on transfer learning performs better on two small datasets, Twitter I and Emotion ROI, which demonstrate that transfer learning can improve accuracy and reduce the risk of overfitting on small datasets.

4.2.2 Comparative Experiment

To better show the effectiveness of this study, the proposed method is compared with the current mainstream methods. Table 9 shows the results of different methods on the three challenging datasets, FI, Twitter I, and Emotion ROI. This study designs a multilayer attentional interaction module and

an adaptive feature aggregation module, which not only enhances the discriminative features of the images but also aggregates the semantics and details of different layers. The experimental results show that this model is more efficient than the other methods in the table.

Table 8: Results without transfer learning

Databases	Methods	Accuracy	Recall
Twitter I	Normal	84.66	86.85
	Without-transfer learning	83.78	84.86
Emotion ROI	Normal	84.96	86.76
	Without-transfer learning	83.24	84.98

Table 9: The comparison between our method and other methods

Databases	Methods	Accuracy	Recall
FI	Rao et al. [34]	62.79	68.90
FI	ResNet101 [35]	75.76	82.63
FI	AR [13]	86.35	87.63
FI	Rao et al. [10]	87.51	92.85
FI	Qu et al. [19]	89.86	94.29
FI	Ours	**90.27**	**94.36**
Twitter I	PCNN [32]	76.36	79.56
Twitter I	AR [13]	81.06	83.45
Twitter I	Wu et al. [14]	81.65	85.79
Twitter I	Qu et al. [19]	83.88	86.68
Twitter I	Ours	**84.66**	**86.85**
Emotion ROI	PCNN [32]	74.06	78.46
Emotion ROI	AR [13]	81.26	83.54
Emotion ROI	Wu et al. [14]	83.04	84.75
Emotion ROI	Qu et al. [19]	84.19	85.94
Emotion ROI	Ours	**84.96**	**86.76**

As shown in Table 9, the performance of models based on deep learning far exceeds that of handcrafted models. Then, compared with ResNet101 model and PCNN model that only use global features, AR model, Wu L's model, Rao T's model, Qu's model and our proposed model can achieve classification accuracy over 80% on the three datasets. The reason is that these models assist in image emotion classification by extracting features from the emotional region of the image.

Among the five models which combine emotional region features and global features for emotion classification, Qu's model and our proposed model are significantly better than the other models. Both of these models obtain higher-layer semantic information and lower-layer high resolution detail information through attention and multi-layer features fusion, so as to predict more accurate emotional regions. Compared with Qu's model, the accuracy of our model is further improved,

reaching 90.27 on the FI dataset and closing to 85 on both smaller datasets. This is attributed to the multilayer attentional interaction module and the adaptive feature aggregation module. The multilayer attentional interaction module calculates the attention feature map many times on the basis of layer-by-layer fusion which strengthens important features and enables higher-layer features to guide lower-layer features learning. The adaptive feature aggregation module adaptively generates a set of aggregated weights through two fully connected layers, enhancing the original features of each layer. Eventually, the semantics and details of the different layers are aggregated through gated fusion and residual connections to complement the missing information. In addition, the excellent performance of the model on small datasets also proves the role of transfer learning. In summary, the above analysis and classification results demonstrate the superiority of this proposed model.

5 Conclusion

This study proposes an image emotion classification network based on multilayer attentional interaction and adaptive feature aggregation. Different from the existing methods that rely on singlelayer features to find emotional regions, this study designed a multilayer attention interaction module. This module calculates spatial attention maps for higher-layer semantic features and fusion features through multilayer shuffle attention module. Through layer-by-layer up-sampling and gating operations, the higher-layer features guides the lower-layer features to learn, eventually achieving sentiment region prediction at the optimal scale. To complement the important information lost by layer-by-layer fusion, this study designs an adaptive feature aggregation module. The module firstly uses global average pooling to compress spatial information and connect channel information from all layers. Then, the module adaptively generates a set of aggregated weights through two fully connected layers to augment the original features of each layer. Eventually, the semantics and details of the different layers are aggregated through gating operations and residual connectivity to complement the lost information. To reduce overfitting on small datasets, the network is pre-trained on the FI dataset, and further weight fine-tuning is performed on the small dataset. The experimental results on the FI, Twitter I and Emotion ROI datasets show that the proposed network exceeds existing image emotion classification methods, with accuracies of 90.27%, 84.66% and 84.96%.

Further, we will study in the following directions: (i) an image may inspire multiple emotions at the same time, so we try to use multi-label learning [36] into image emotion classification to achieve more accurate emotion prediction. (ii) we will explore the association of text emotion and image emotion through attention mechanisms to achieve more accurate multimodal emotion classification.

Acknowledgement: We are grateful to Nanjing University of Information Science and Technology for providing study environment and computing equipment.

Funding Statement: This study was supported, in part, by the National Nature Science Foundation of China under Grant 62272236; in part, by the Natural Science Foundation of Jiangsu Province under Grant BK20201136, BK20191401.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] D. Joshi, R. Datta, E. Fedorovskaya and Q. T. Luong, "Aesthetics and emotions in images," *IEEE Signal Processing Magazine*, vol. 14, no. 5, pp. 94–115, 2011.

[2] T. Rao, X. X. Li, H. M. Zhang and M. Xu, "Multi-level region-based convolutional neural network for image emotion classification," *Neurocomputing*, vol. 333, no. 6, pp. 429–439, 2019.
[3] S. Corchs, E. Fersini and F. Gasparini, "Ensemble learning on visual and textual data for social image enotion classification," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 4, pp. 1–14, 2019.
[4] Y. W. He and G. G. Ding, "Deep transfer learning for image emotion analysis: Reducing marginal and joint distribution discrepancies together," *Neural Processing Letters*, vol. 51, no. 5, pp. 2077–2086, 2020.
[5] X. X. Yao, S. C. Zhao, Y. K. Lai, D. Y. She, J. Liang *et al.,* "APSE: Attention-aware polarity sensitive embedding for affective image retrieval," *IEEE Transactions on Multimedia*, vol. 23, no. 1, pp. 4469–4482, 2020.
[6] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. of the 18th ACM Int. Conf. on Multimedia*, Firenze, Italy, pp. 83–92, 2010.
[7] S. C. Zhao, Y. Gao, X. L. Jiang, H. X. Yao, T. S. Chua *et al.,* "Exploring principles-of-art features for image emotion recognition," in *Proc. of the 22nd ACM Int. Conf. on Multimedia*, Orlando, Florida, USA, pp. 47–56, 2014.
[8] Q. You, H. L. Jin and J. B. Luo, "Visual sentiment analysis by attending on local image regions," in *Proc. of the Thirty-First AAAI Conf. on Artifificial Intelligence*, San Francisco, CA, USA, vol. 31, no. 1, pp. 4–9, 2017.
[9] S. J. Fan, Z. Q. Shen, M. Jiang and B. Koenig, "Emotional attention: A study of image sentiment and visual attention," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2018)*, Salt Lake City, UT, USA, 2018.
[10] T. Rao, X. X. Li and M. Xu, "Learning multi-level deep representations for image emotion classification," *Neural Processing Letters*, vol. 51, no. 3, pp. 2043–2061, 2020.
[11] K. K. Song, T. Yao, Q. Ling and T. Mei, "Boosting image sentiment analysis with visual attention," *Neurocomputing*, vol. 312, no. 6, pp. 218–228, 2018.
[12] Z. Li, C. Y. Lang, J. H. Liew, Y. D. Li, Q. B. Hou *et al.,* "Cross-layer feature pyramid network for salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 99, 2021.
[13] J. F. Yang, D. Y. She, M. Sun and M. M. Cheng, "Visual sentiment prediction based on automatic discovery of affective regions," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2513–2525, 2018.
[14] L. F. Wu, M. C. Qi, M. Jian and H. Zhang, "Visual sentiment analysis by combining global and local information," *Neural Processing Letters*, vol. 51, no. 3, pp. 1–13, 2019.
[15] W. Sun, G. Z. Dai, X. R. Zhang, X. Z. He and X. Chen, "TBE-Net: A three-branch embedding network with part-aware ability and feature complementary learning for vehicle re-identification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 14557–14569, 2022.
[16] J. Chen, Z. Zhou, Z. Pan and C. Yang, "Instance retrieval using region of interest based cnn features," *Journal of New Media*, vol. 1, no. 2, pp. 87–99, 2019.
[17] X. X. Yao, D. Y. She, S. C. Zhao, J. Liang, Y. K. Lai *et al.,* "Attention-aware polarity sensitive embedding for affective image retrieval," in *Proc. of the ICCV*, Seoul, Korea, pp. 1140–1150, 2019.
[18] D. T. Priya and J. D. Udayan, "Affective emotion classification using feature vector of image based on visual concepts," *International Journal of Electrical Engineering Education*, vol. 57, no. 3, pp. 1–22, 2020.
[19] H. C. Qu, C. M. Qing, X. M. Xu and J. X. Jin, "Multi-level context pyramid network for visual sentiment analysis," *Sensors*, vol. 21, no. 6, pp. 2136–2155, 2021.
[20] K. C. Peng, A. Sadovnik, A. Gallagher and T. Chen, "Where do emotions come from? Predicting the emotion stimuli map," in *Proc. of the 2016 IEEE Int. Conf. on Image Processing (ICIP)*, Phoenix, AZ, USA, 2016.
[21] K. Sun, B. Xiao, D. Liu and J. D. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, vol. 43, pp. 5693–5703, 2019.
[22] J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks," *Proc. of the IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 7132–7141, 2018.

[23] Q. L. Wang, B. G. Wu, P. F. Zhu, P. H. Li, W. M. Zuo *et al.,* "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 2575–7075, 2020.
[24] S. Woo, J. Park, J. Y. Lee and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. of the ECCV*, Munich, MUC, Germany, pp. 3–19, 2018.
[25] Q. L. Zhang and Y. B. Yang, "Shuffle attention for deep convolutional neural networks," in *Proc. of the ICASSP*, Toronto, Ontario, Canada, pp. 7132–7141, 2021.
[26] F. Z. Zhuang, Z. Y. Qi, K. Y. Duan, D. B. Xi, Y. C. Zhu *et al.,* "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2021.
[27] X. C. Wei, H. F. Lin, L. Yang and Y. H. Yu, "A convolution-LSTM-based deep neural network for cross-domain MOOC forum post classification," *Information*, vol. 8, no. 3, pp. 93, 2017.
[28] X. R. Zhang, J. Zhou, W. Sun and S. K. Jha, "A lightweight CNN based on transfer learning for COVID-19 diagnosis," *Computers, Materials & Continua*, vol. 72, no. 1, pp. 1123–1137, 2022.
[29] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," *Proc. of the 28th International Conference on Neural Information Processing Systems*, vol. 2, pp. 3079–3087, 2015.
[30] S. J. Fan, Z. Q. Shen, M. Jiang, B. L. Koenig, J. Xu *et al.,* "Emotional attention: A study of image sentiment and visual attention," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 7521–7531, 2018.
[31] Q. You, J. Luo, H. Jin and J. C. Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark," in *Proc. of the 30th AAAI Conf. on Artifificial Intelligence*, Phoenix, AZ, USA, vol. 30, 2016.
[32] Q. You, J. Luo, H. Jin and J. C. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. of the 29th AAAI Conf. on Artifificial Intelligence*, Austin, Texas, USA, vol. 29, no. 1, pp. 381–388, 2015.
[33] S. C. Zhao, X. X. Yao, J. F. Yang, G. L. Jia, G. G. Ding *et al.,* "Affective image content analysis: Two decades review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6729–6751, 2021.
[34] T. Rao, M. Xu, H. Y. Liu, J. P. Wang and I. Burnett, "Multi-scale blocks based image emotion classifification using multiple instance learning," in *Proc. of the 2016 IEEE Int. Conf. on Image Processing (ICIP)*, Phoenix, AZ, USA, pp. 634–638, 2016.
[35] K. M. He, X. Y. Zhang, S. Q. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770–778, 2016.
[36] J. He, C. Wang, H. Wu, L. Yan and C. Lu, "Multi-label chinese comments categorization: Comparison of multi-label learning algorithms," *Journal of New Media*, vol. 1, no. 2, pp. 51–61, 2019.