CGLF-Net: Image Emotion Recognition Network by Combining Global Self-Attention Features and Local Multiscale Features

Yutong Luo, Xinyue Zhong, Minchen Zeng, Jialan Xie, Shiyuan Wang, and Guangyuan Liu, Jr., *Member, IEEE*

***Abstract*— Convolutional neural networks (CNNs) are commonly employed for image emotion recognition owing to their ability to extract local features; however, they have difficulty capturing the global representations of images. In contrast, self-attention modules in a visual transformer network can capture long-range dependencies as global features. Some studies have shown that an image's local and global features determine the emotions of the image and that some local regions can generate an emotional prioritization effect. Therefore, we proposed combining global self-attention features and a local multiscale features network (CGLF-Net) to recognize an image's emotion, extracting image features from global and local perspectives. Specifically, the cross-scale transformer network is employed instead of convolution operations in the global feature branch to enhance its model feature representation. In the local feature branch, the improved feature pyramid module is applied to extract features from different sensory fields, thereby combining semantic information with different scales. Furthermore, the local attention module based on class activation maps guides the network to focus on locally salient regions. In addition, using multibranch loss functions, local and global feature branches are combined to enhance the ability to capture a comprehensive set of features. Consequently, the proposed network achieves recognition accuracies of 75.61% and 65.01% on the FI-8 benchmark dataset and Emotion-6 benchmark dataset, respectively. These results show that the proposed CGLF-Net reliably address the difficulty of extracting global features using CNNs, representing the classification performance of the state-of-the-art.**

***Index Terms*—class activation map, global and local features, image emotion recognition, local attention, multibranch loss function, multiscale, self-attention.**

I. Introduction

WITH the development of social media, a growing number of people are expressing their emotions by sharing pictures on the internet, which has led to a research focus on image emotion recognition. Psychological studies have shown that visual stimuli can influence human emotions [1-4]. It is important to enable machines to recognize and understand emotions in imagery for image retrieval [5, 6], opinion mining [7-10], and aesthetic classification [11].

Unlike traditional image classification and segmentation tasks, image emotion recognition is challenging because it considers people's reactions to visual content. Although the various features of an image may arouse different emotions, the complexity of the content increases the difficulty of image emotion recognition [12]. Moreover, images with a multitude of content may express multiple similar emotions, implying a large intraclass gap between two images of the same category [13, 14].

Early image emotion recognition studies attempted to perform emotion recognition using low-level visual features, such as image color and texture [15]. However, it is difficult to accurately relate these features to emotions, and manually designed features have considerable limitations. Therefore, with the development of deep learning, many studies have begun using related models for image emotion recognition. One common method is the use of convolutional neural networks (CNNs) to extract image features [16]. Relevant experimental results have shown that end-to-end classification methods that rely on CNNs are more effective than manually designed feature classification methods in image emotion recognition tasks [17].

In early studies, each region of an image was treated equally during learning, with different features extracted for different tasks. Subsequently, psychological studies showed that different regions and varied content contribute differently to emotional arousal [18]. Some local regions in an image may contain more discriminative information regarding emotions than other parts. Zhao et al. [19] found that using both local features and global features improves results, suggesting that image emotion recognition should combine local and global feature extraction to obtain more discriminative representations.

Consequently, obtaining appropriate features from images is essential to image emotion recognition. In previous studies, single visual features were considered global representations; thus, they disregarded the emotional responses evoked by local areas of an image [13, 20]. Some recent studies have attempted to combine local and global features using CNNs to extract

The authors thank all the researchers involved in the experiment. This work was supported by the National Natural Science Foundation of China (Nos. 61872301 and 61472330). Corresponding Author: GuangYuan Liu.

Y. Luo, X Zhong, M. Zeng, J. Xie, S. Wang and G. Liu are with t he School of Electronic and Information Engineering, Southwest China University, Chongqing, CO 400715 China. (email: lyt252012778@email.swu.edu.cn; xzhong3@utas.edu.au; zmczmc@email.swu.edu.cn; jialanxie@email.swu.edu.cn; wangshiyuan@e-mail.swu.edu.cn; liugy@swu.edu.cn).

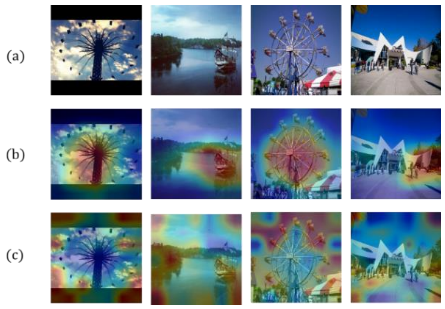


Fig. 1 Images labeled as "sad:" (a) Original images, (b) ResNet-50 attention map, (c) Swin-based network attention map.

global features [21]. However, it is difficult to obtain global representations because convolutional operations are effective only at extracting local feature information; thus, the extracted features cannot adequately summarize the global information [22-24].

Recently, Vaswani et al. [25] proposed a transformer structure instead of a convolution operation for global image feature extraction. The Vision Transformer (ViT) method divides an image into a fixed number of patches with location information to construct a sequence of tokens and extract the parameter vector as a visual representation. Furthermore, sophisticated spatial transformations and long-distance feature reliance are adopted as global representations via a self-attention mechanism and multilayer perceptron (MLP) structure. Liu et al. [26] subsequently proposed the Swin Transformer algorithm, which applies a pyramid structure window attention method and establishes a cross-window relationship using shifted windows. Thus far, the transformer structure has been widely utilized in image classification tasks and has achieved encouraging results by obtaining global representations from images for classification. Fig. 1 shows the attention graphs of traditional CNNs and transformer networks. CNNs focus on local image regions during classification, whereas transformer networks weigh multiple regions. Moreover, research in image emotion recognition using the combination of convolutional and self-attention features to extract local and global features is not available. Consequently, owing to the excellent global feature extraction capability of transformers, this study attempts to combine the advantages of transformer networks with CNNs to enhance image sentiment recognition performance, which will explore its feasibility with regard to image emotion recognition.

To combine the local and global information of images, this study proposes a two-branch network structure, which is referred to as the combining global self-attention features and local multiscale features network (CGLF-Net), which enhances the network's representation of image sentiment as a function containing both local features and global features. The network consists of a local feature branch based on a CNN structure and a global feature branch based on a transformer network. The local feature branch obtains local features that are highly correlated with sentiment labels to avoid the influence of noise in nonsentiment regions. In the local feature branch, the class activation technique is applied to guide the extraction of local features. Specifically, global feature branching combines the information of different sentiment regions to prevent the feature information from being excessively homogeneous. A new loss function is also proposed to render the output distributions of the two branches consistent, and a class activation map (CAM) loss term is added to the loss function to improve accuracy.

The main contributions of this work are summarized as follows:

1) We propose a two-branch network named CGLF-Net. In the local feature branch, we extract local multiscale features of an image via a modified feature pyramid network and CAM weighted local attention module (CW-LAM). Moreover, we obtain global self-attention features by a cross-scale transformer network in the global feature extraction branch. The results indicate that the combination of global self-attention features and local multiscale features can have a complementary effect, resulting in richer emotional features.

2) Kullback–Leibler divergence is introduced to the loss function to maintain consistency in the output distributions of both network branches. A CAM loss term is also added to the loss function to further improve performance. The results of the ablation experiments show that the emotion recognition rate of the two-branch network can be improved by the loss function we designed.

3) The proposed CGLF-Net for image emotion recognition is experimentally validated by the FI-8, Emotion-6 and WEBEmo datasets. The results show that the network achieved promising performance compared to state-of-the-art models. In particular, the accuracy on Emotion-6 was significantly improved by 5% compared to existing methods, which reached the highest level.

II. Related Work

This study focuses on the image emotion recognition problem, and the approach used is highly relevant to transformer networks. In this section, we review existing methods in terms of the two aspects mentioned above.

A. Image Emotion Recognition

Image emotion recognition, which is also referred to as emotional content analysis, is an important field of artificial intelligence research that is aimed at recognizing or generating emotions such as those generated by humans after viewing images. Currently, the main categories of image emotion recognition are categorical emotion states (CESs) [27-30] and dimensional emotion spaces (DESs) [31, 32]. DES models provide flexible descriptions of emotions by defining them in 2D, 3D, or higher-dimensional Cal systems in continuous space, and typical DES models include the valence–arousal space [33]. In contrast, in the CES model, emotions are directly defined as several discrete feelings, such as anger, fear, and sadness. Compared with the DES model, the CES model is more intuitive; therefore, most studies use the CES model. The CES model divides emotions into two categories—positive and negative—using strict psychological theory. Ekman [28] further classified emotions into six categories (i.e., anger,

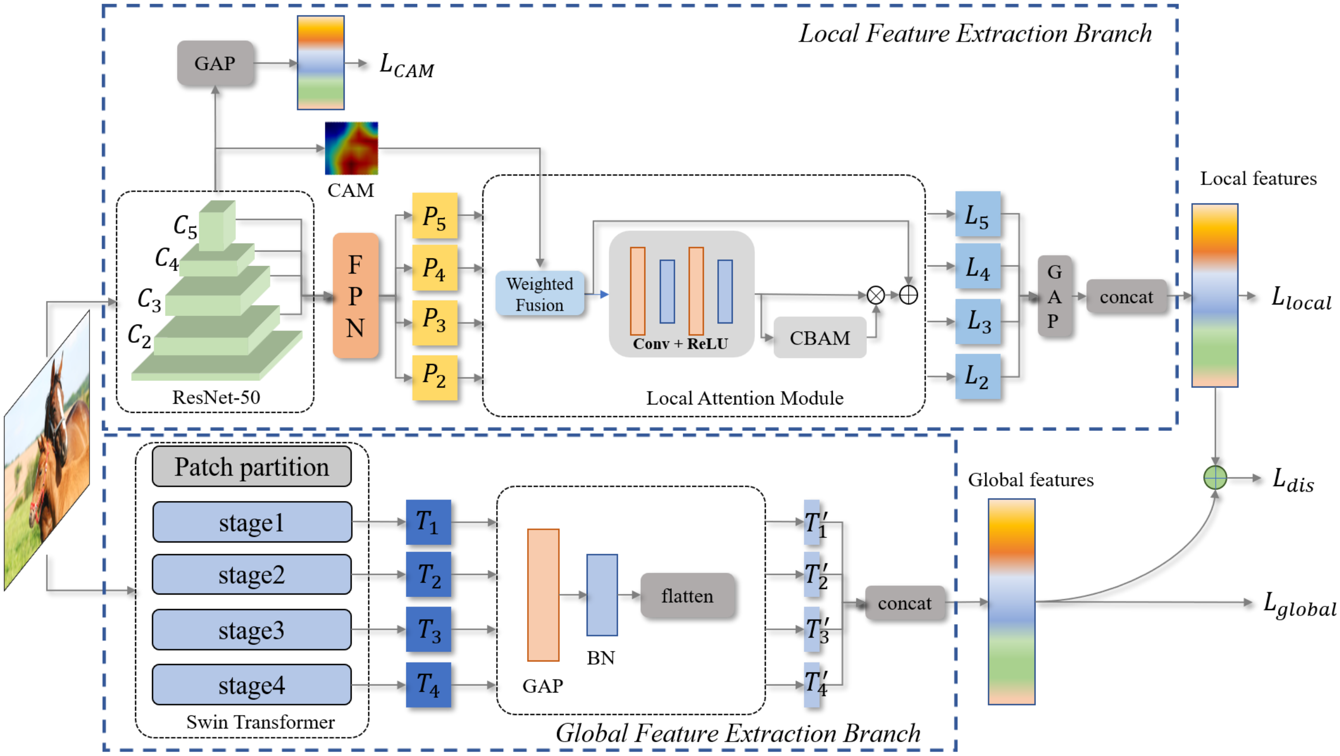


Fig. 2. Structure of CGLF-Net, consisting of two branches: (a) local features extraction branch and (b) global features extraction branch. The entire network was trained using a multitask loss function.

disgust, fear, happiness, sadness, and surprise), whereas Mikel et al. [30] classified emotions into four positive emotions (i.e., amusement, awe, contentment, and excitement) and four negative emotions (i.e., anger, disgust, fear, and sadness). To allow for cross-sectional comparisons, this study is based primarily on the CES model.

In previous studies, different combinations of low-level visual features were extracted for emotion recognition, with color and texture being the most representative. Lee et al. [32] proposed MPEG-7, a complex combination of features consisting of color and texture. Machajdik et al. [34] used features extracted from aesthetic art concepts, which included color, texture, composition, and content. Based on primary visual features, such as color and texture, some studies have attempted to obtain more complex features using histograms [35, 36]. Because low-level features can only provide basic and local information in an image, the features lack the representation of semantic information in the image. Furthermore, low-level features are susceptible to interference from environmental factors such as lighting and noise, which leads to a decrease in emotion recognition accuracy.

Because the relationship between low-level visual features and image emotion is difficult to understand, some studies have extracted mid-level features to bridge the gap between low-level visual features and high-level emotions. Mid-level image features primarily include attributes, sentributes, constructs, and compositions. Yuan et al. [37] proposed Sentribute, an image emotion recognition algorithm that uses detected facial expression features to determine emotional polarity. The features designed by Wang et al. [38] provide a map comparison and have better interpretability and comprehensibility than traditional features. Rao et al. [39] found that the use of bag-of-visual words can effectively model the emotional information contained in local regions of an image while avoiding false recognition caused by extracting features from the entire image. According to different artistic principles, Zhao et al. [12] proposed principles-of-art-based emotion features, a comprehensive representation based on balance, emphasis, harmony, change, hierarchy, and movement. Lu et al. [40] proposed three 1D visual feature scalars (i.e., roundness, angularity, and visual complexity) that were effective for practical image emotion recognition. Although mid-level features contain some of the abstract information in images, they still cannot accurately represent the complexity and diversity of emotions. Simultaneously, the associations and patterns among features are difficult to interpret and understand, thus limiting the effectiveness and reliability of emotion recognition. There are certain differences in the performance of intermediate features in different images, leading to a low accuracy of emotion recognition.

High-level image features refer to semantic information, which is easy to understand. These features are more closely related to image emotions than low- and mid-level features. Borth et al. [41] developed SentiBank, a large visual emotion ontology consisting of 1,200 adjective–noun pairs (ANPs), by combining psychology and taxonomy (folksonomy) features extracted from social multimedia. Based on SentiBank, Jou et al. [42] proposed a large-scale multilingual visual emotion

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3289762

ontology. To bridge the emotional gap between image content and evoked emotions, Ali et al. [43] introduced high-level concepts (HLCs), which contain objects and places as features for image emotion recognition. The diversity and complexity of semantic information in images makes it difficult to accurately describe images based on hand-designed high-level features. Additionally, for some images, such as abstract paintings, features may not be sufficient to serve as emotional representations.

With advancements in deep learning, CNNs have achieved breakthroughs in various fields. Learning-based features have shown outstanding performance in affective image content analysis (AICA), and CNN-based methods have been applied to image emotion recognition tasks. Peng et al. [13] were the first researchers to apply CNN models to image emotion recognition tasks, and their experimental results demonstrated that the CNN-based approach outperformed the hand-designed feature approach. Chen et al. [44] trained the classification of 1,200 ANPs using a CNN based on SentiBank, and the resulting deep model, DeepSentiBank, outperformed its predecessor. You et al. [45] used a progressive CNN pretrained on 500,000 weakly labeled images from SentiBank to obtain a pretrained model, discovering that the trained model was robust against small-scale labeled datasets. Chen et al. [21] employed a pretrained CNN and pretrained AlexNet to obtain image features for image emotion recognition. Rao et al. [39] utilized a multilevel network structure to fuse low-level features obtained from different CNN learning layers with high-level features to achieve image emotion recognition.

In these previous studies, each region of an image was treated equally, and different global features were extracted for different tasks. According to psychological theory, emotional content is associated with certain important regions in an image; therefore, an increasing number of studies have investigated ways to extract local features to enhance image emotion recognition. Rao et al. [46] proposed the extraction of a multilevel depth representation using a feature pyramid network (FPN), which generated suitable local regions using emotion region suggestion methods to remove redundant nonemotion regions. You et al. [47] used an attention model to identify local areas that triggered certain emotions in observers and extracted their features to improve visual emotion classification. Moreover, Yang et al. [48] used a pretrained target detection tool to extract targets in images and remove the influences of redundant information on classification. Zhao et al. [49] proposed a CNN framework that contains detection and classification branches. In the detection branch, soft emotion maps are generated by classifying features, and global features are combined with soft emotion maps to obtain comprehensive local feature information. She et al. [50] introduced a weakly supervised coupled convolutional network (WSCNet) dedicated to the automatic selection of relevant weak annotations when provided, thereby significantly reducing the annotation burden. Zhang et al. [51] proposed an end-to-end deep neural network for image emotion recognition using emotion intensity learning and succeeded by leveraging emotion intensity maps as additional supervisory information for the network. Most recent studies have designed networks for correlating image emotions with local image features and directly applied convolutional neural networks to extract global features of images. However, CNNs are more adept at extracting local features of images, and the global information of images cannot be well characterized using only CNNs.

B. Vision Transformers (ViTs)

Transformer is a deep learning model that is widely utilized in many fields, such as natural language processing and computer vision. Transformer networks were first used as a sequence-to-sequence model for machine translation, and later, Dosovitskity et al. [25] proposed the vision transformer (ViT) network, which favored a pure transformer network instead of the traditional CNN structure for direct image classification. ViT networks decompose images into sequences of tokens with fixed lengths and relationally model these token sequences through multiple transformer layers. ViT networks indicate that pure Transformer networks are very effective in computer vision tasks while having significant advantages over CNN networks on large-scale datasets [25]. However, ViT networks require the computation of global attention of images, resulting in network models that must be pretrained on large-scale datasets (e.g., ImageNet-22K, JFT-300 M, etc.) to be comparable to CNN networks. Additionally, ViT networks suffer from excessive computational overhead for high-resolution image inputs [52].

To address the problem that ViT networks are difficult and ineffective to train on small-scale datasets, DeiT [52] proposes a training strategy on smaller datasets (e.g., ImageNet1K) that focuses on introducing distillation means and setting superior hyperparameters during the training process. The experimental results show that the training strategy used with DeiT makes the ViT model outperform the CNN network on small-scale datasets. The computational complexity of the ViT network is quadratic to the image size, resulting in a high computational overhead of the model. To address this problem, Wang et al. [53] proposed the pyramid vision transformer network (PVT), which introduces a pyramid structure in the ViT network, gradually decreases the resolution of the feature map, and increases the number of feature channels as the network deepens. The experimental results show that the PVT network has lower model parameters and higher classification accuracy. Liu et al. [26] proposed the Swin Transformer network, which replaces the fixed window segmentation in the original ViT network with sliding window segmentation, thus achieving a balance of speed and accuracy in image classification.

III. METHODOLOGY

The purpose of this study is to improve image emotion recognition by combining local image features with global features. Unlike previous studies in which CNNs were used to extract global image features, the proposed method used transformer networks instead of CNNs as the image global feature extraction method. The network structure is shown in Fig. 2. The

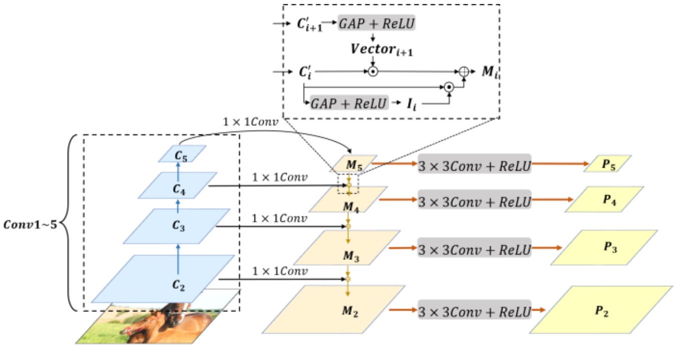


Fig. 3. Feature pyramid network structure. c_2–c_5 = Convolutional blocks 2–5; *Conv* = Convolution; ReLU = Activation layer; GAP = Global average pooling; M_2–M_5 = Middle features 2–5; P_2–P_5 = Pooling layers 2–5.

network comprises local and global feature extraction branches.

A. Local Feature Extraction Branch

As mentioned in Section 1, the emotion of an image is highly correlated with local features. In the local feature extraction branch, this study uses a CNN as the base stem network to collect local features in layers using convolutional operations and retain local cues as features. CNNs capture local feature information more effectively than the self-attention modules of a transformer.

1) **Multiscale Module**

In the local feature extraction branch, this study uses a CNN as the base stem network to collect local features in layers using convolutional operations and to retain local cues as features. CNNs capture local feature information more effectively than the self-attention modules of a transformer.

Previous studies have demonstrated that the use of multiscale features can significantly improve image emotion recognition performance [39, 54]. However, in the original feature pyramid network, each scale feature has different resolution and semantic information, and direct fusion will lead to the loss of the original semantic information. To improve the performance of multiscale information extraction, this study uses an improved FPN to extract multiscale image information. The detailed structure, which is shown in Fig. 3, consists of three parts (i.e., bottom-up path, top-down path, and lateral connection) and generates features at different scales through the bottom-up paths, where features generated at higher levels have smaller scales [55]. The features generated by extracting the bottleneck layers of the four convolutional blocks in the CNN serve as feature pyramid inputs $\{C_2, C_3, C_4, C_5\}$. Owing to the large feature scale generated by the first convolutional layer, it is omitted to reduce network overhead.

In the top-down process, the feature fusion process is improved to reduce the semantic information loss of features at each scale in the feature pyramid network. The specific implementation process is described as follows: First, the top-level feature $C_{i+1} \in \mathbb{R}^{H\times H\times D}$ is feature-dimensioned by a 1 × 1 convolution to obtain the feature $M_{i+1} \in \mathbb{R}^{H\times H\times D'}$, where D and D' indicate the number of channels. The number of channels of this feature is aligned with the number of channels of the next layer of features. Subsequently, the semantic vector $Vector_{i+1} \in \mathbb{R}^{1\times 1\times D'}$ is extracted for the feature M_{i+1}, which represents the key semantic information in the feature. The feature $S_i \in \mathbb{R}^{H\times H\times D'}$ is obtained by feature fusion on feature C_i using the semantic vector $Vector_{i+1}$. To maintain the semantic information in the original features, we extract the key information vector $I_i \in \mathbb{R}^{1\times 1\times D'}$ for the features and use this vector for semantic reinforcement of the original features C_i. Feature S_i is fused with feature C_i for feature fusion. The overall computational process is presented as follows:

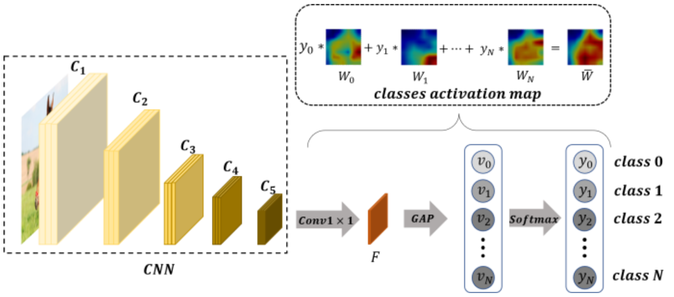


Fig. 4. Structure of image emotion class activation graph subnetwork. W_1–W_N = Activation maps 1–*N;* $\bar{W}$ = *Local attention map*; c_1–c_5 = Convolutional blocks 1–5; *CNN* = Convolutional neural network; *Conv* = Convolution; *GAP* = Global average pooling; v_1–v_N = Pooling layers 1–*N*; y_1–y_N = Softmax output 1–*N*.

$$M_i = \begin{cases} Conv_{1\times 1}(C_i), & i = 5 \\ I_i \odot C_i + S_i, & i = 2,3,4 \end{cases}, \tag{1}$$

$$S_i = Vector_{i+1} \odot C_i, \quad i = 2,3,4, \tag{2}$$

$$Vector_{i+1} = \mathrm{Re}LU(\frac{\sum_{m,n}(M_{i+1})_{m,n}}{h^M \times h^M}), \quad i = 2,3,4, \tag{3}$$

$$I_i = \mathrm{Re}LU(\frac{\sum_{x,y}(C_i)_{x,y}}{h^C \times h^C}), \quad i = 2,3,4, \tag{4}$$

where $\odot$ denotes elementwise multiplication, h^M denote the height and width of feature M_{i+1}, m denotes the m^{th} row of feature M_{i+1}, n denotes the n^{th} column of feature M_{i+1}, h^C denote the height and width of feature C_i, x denotes the x^{th} row of feature C_i, and y denotes the y^{th} column of feature C_i.

In the lateral connections, the features M_i at each scale must be downscaled by 1 × 1 convolution so that the number of channels of all multiscale features P_i remains the same. In our experiments, the number of feature channels was set to 128, and the calculation procedure is shown below.

$$P_i = \mathrm{Re}LU(Conv_{1\times 1}(M_i)), \quad i = 2,3,4,5, \tag{5}$$

With the feature pyramid network, the feature maps $\{P_2, P_3, P_4, P_5\}$ fuse features of different scales and semantic strengths to enhance network access to detailed information.

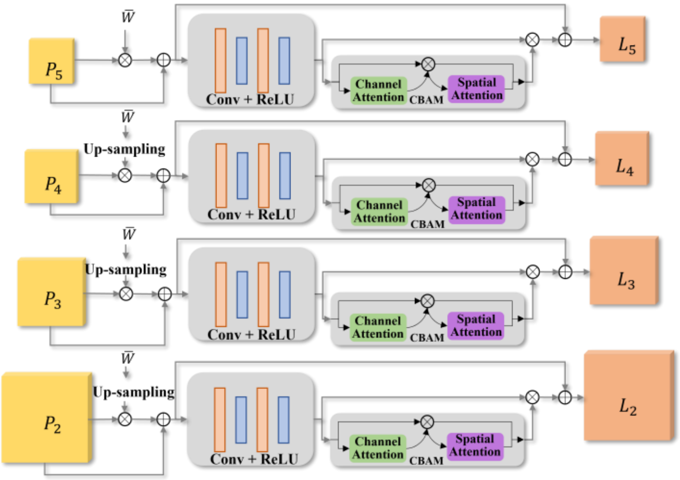


Fig. 5. Structure of the CAMs weighting local attention module (CW-LAM). BN = Batch normalization; CBAM = Convolutional block attention module; *Conv* = Convolution layer; P_2–P_5 = Multiscale features 2–5; L_2–W_5 = Local features 2–5; $\overline{W}$ = *Local attention map*; *Upsampling* = Nearest neighbor interpolation.

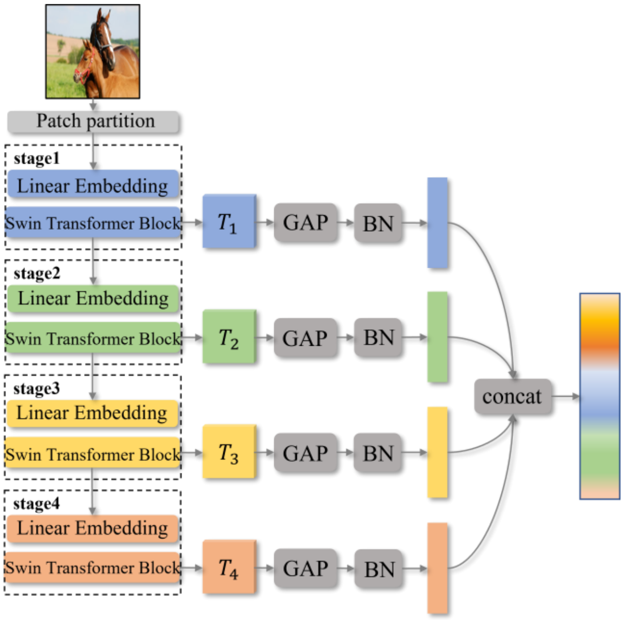


Fig. 6. Structure of the global feature extraction network. The network is divided into four stages, and the output sizes of the stages are $H/4\times W/4\times C$, $H/8\times W/8\times 2C$, $H/16\times W/16\times 4C$, and $H/32\times W/32\times 8C$. GAP = Global average pooling layer; BN = Batch normalization layer; T_1–T_4 = Global features 1-4; H = Height; W = Width.

2) Local Attention Module

In psychological theory, emotional images are defined as human responses to visual inputs [2, 56]. The role of the attention mechanism is to precisely determine salient regions in the image. This study's goal is to increase performance by using attention mechanisms while focusing on important features and suppressing unnecessary features. Thus, the network is allowed to autonomously focus on locally salient features using a local attention mechanism.

To further enhance the network for local feature extraction, most previous studies used a CAM to obtain image-like activation maps to guide the network to extract local features. However, a CAM must be obtained by multiplying the parameters of the fully connected layer with the last layer's feature map, which cannot be directly obtained via forward propagation. In this study, the class activation map (CAM) is directly obtained from the forward propagation of the network, and based on it, the local attention maps of the images are obtained. Eventually, the local attention weights combined with the convolutional block attention module (CBAM) constitute the new local attention module, which is referred to as the class activation graph weighted local attention module (CW-LAM).

The ResNet network is employed as the base stem network, and C_5 features are extracted using one layer of convolution and four convolution blocks. The channels of the $C_5 \in \mathbb{R}^{H\times H\times N}$ features are converted to emotion category numbers using 1×1 convolutions, and the global average pooling (GAP) layer output is mapped to the final emotion category.

$$W_i = \sum_{k=0}^{K-1} A_{k,i} \cdot F_k, \quad i = 0,1,2,...,N, \tag{6}$$

where $A_{k,i}$ denotes the value of the kth row and i-th column in the weight matrix A and F_k denotes the feature matrix of the kth channel in feature F. After obtaining the feature matrix F, it is input to the global average pooling layer to obtain the vector $V = \{v_n\}, n = 0,1,2,\dots,N$. The calculation formula is expressed as follows:

$$v_n = \frac{\sum_{i,j}(F_n)_{i,j}}{H\times H}, \quad n = 0,1,2,\dots,N, \tag{7}$$

where $(F_n)_{i,j}$ denotes the value of the i-th row and j-th column of the n-th channel of the feature map F. The weight vector $V = \{v_n\}, n = 0,1,2,\dots,N$ is input to the softmax layer to obtain the final classification weight $P = \{p_n\}, n = 0,1,2,\dots,N$. To obtain the local attention weight $\overline{W}$ of the image, we weight and fuse all class activation maps of the image and choose the classification weight P as the weight of the weighted fusion. The calculation procedure is expressed as follows:

$$\overline{W} = \sum_{i=0}^{N} y_i \times W_i, \tag{8}$$

The local attention map $\overline{W}$ is used to obtain the significant regions in the image that are related to the emotion. The value in the local attention map $\overline{W}$ represents the degree of relevance of a region to the emotion. The local attention map $\overline{W}$ provides local information that can be applied to improve the performance of emotion recognition. To ensure that the model can be properly trained, we normalize the values of the local attention map to [0, 1].

After obtaining the image local attention map $\overline{W}$, the size of the local attention map is changed to the corresponding size of each multiscale feature extracted by the FPN via upsampling. Here, nearest neighbor interpolation was selected as the upsampling method. The image local attention map is weighted

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3289762

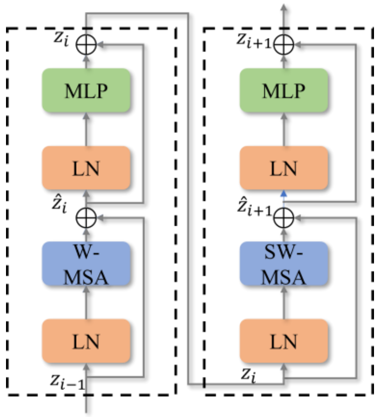


Fig. 7. Structure of two consecutive swin-transformer blocks. LN = Layer normalization layer; MLP = Multilayer perceptron; MSA = Self-attention module; SW-MSA = Shifted-window MSA; W-MSA = Regular window MSA; $\hat{Z}$ = MSA output.

and fused with features of the same size, and the calculation procedure is expressed as follows:

$$P_i' = \bar{W} \odot P_i + P_i, \quad i = 2,3,4,5, \qquad (9)$$

where P_i denotes the i-th multiscale feature map and P_i' denotes the i-th scale feature map after local attention weighting. The number of channels of all multiscale features is set to 128. To further combine the channel information and spatial information of local features, a CBAM is used to refine the local features in this study. A CBAM is a lightweight general-purpose module that can be applied to any CNN framework while disregarding module overhead [57]. CBAM separately computes attention maps in both spatial dimensions and channel dimensions and multiplies them with the feature maps for adaptive feature refinement. The size of the CBAM feature map is the same as that of the original feature map.

In the proposed network structure, the input of the attention module is the multiscale feature P_i', the CBAM output is the channel attention map $M_c \in \mathbb{R}^{1\times1\times C}$, and the spatial attention map is $M_s \in \mathbb{R}^{H\times W\times 1}$. Ultimately, the attention graph corresponding to each scale feature is defined as:

$$Attn_i = M_s(F_i) \odot (M_c(F_i) \odot F_i), \quad i = 2,3,4,5 \, , \qquad (10)$$

where $\odot$ denotes elementwise multiplication. Channel attention information is propagated along the spatial dimension by multiplying the channel attention with the original feature map. The result is then multiplied by spatial attention so that the spatial attention values are propagated along the channel dimension. *Attn*$_i$ represents the final fusion output results corresponding to features at each scale.

The input of the CW-LAM comprises four scales of features extracted by the FPN structure. Their sizes are 7×7×128, 14×14×128, 28×28×128, and 56×56×128, and the output size of the local attention module is the same as the input. The four outputs of the CW-LAM are passed through the GAP layer to obtain a 128-dimensional feature vector, and the four local feature vectors are concatenated to form a 512-dimensional feature vector. This feature vector is fed into the fully connected layer to achieve emotion recognition.

B. Global Feature Extraction Branch

In the field of deep learning, CNNs can be considered hierarchical sets of local features with different perceptual domains [22]. Therefore, they are efficient at extracting local features but lack the ability to acquire global features [23]. In contrast, the Vision Transformer network constructs a sequence of markers by segmenting an image into small blocks with positional embeddings and by extracting the parameterized vectors as a visual representation through the transformer module. After the self-attention mechanism and MLP structure are employed, the ViT networks obtain complex spatial transformations and long-range feature dependencies; therefore, these networks have excellent global feature representation capabilities.

Furthermore, Liu et al. [26] proposed a Swin Transformer network based on ViT, which increases the efficiency of the conventional Swin Transformer network by restricting self-attention to local windows and allowing cross-window connections. Moreover, a hierarchical network structure is employed, resulting in the ability to model at multiple scales. In this study, a Swin Transformer serves as the base stem network for the global feature extraction branch. To retain more underlying information in the global features, we design a cross-scale feature fusion module for emotion recognition by extracting features from different stages in the Swin Transformer. The structure of the global feature extraction network is shown in Figure 6.

The Swin Transformer network uses the patch partition module to split the input red–green–blue channel image into multiple nonoverlapping patches. In this study, the size of each patch is set to 4 × 4 so that the feature dimension of each patch is 4 × 4 × 3. Therefore, after segmentation, the original image of size *H* × *W* × 3 is $H/4 \times W/4 \times 48$.

The linear embedding and Swin Transformer blocks are combined into one stage, and the entire Swin Transformer network contains four stages. First, the patch is mapped to an arbitrary dimension, *C*, after passing through the linear embedding module, where the obtained feature size is $H/4 \times W/4 \times C$. Subsequently, each patch is fed to the Swin Transformer block to calculate the self-attention value.

The Swin Transformer block includes a moving-window-based multihead attention module as well as an MLP module. LayerNorm layers are applied before each multihead self-attentive (MSA) mechanism and MLP for normalization, and a residual module is attached after each.

The multiheaded attention module is shown in Figure 7. In the multiheaded attention module, the multihead self-attention is calculated similarly to the ViT algorithm but with the addition of a relative position bias to the self-attention. The calculation formula is presented as follows:

$$Attention(Q,K,V) = SoftMax(QK^T / \sqrt{d} + B)V, \qquad (11)$$

where $Q, K, V \in \mathbb{R}^{g^2 \times d}$ include the query, key, and value matrices, respectively; d is the query/key dimension; and g^2 is the number of patches in a window. Because the relative position along each axis lies in the range [-g+1, g-1], a smaller-sized bias matrix, $\hat{B} \in \mathbb{R}^{(2g-1)\times(2g-1)}$, is parameterized, and values in B are obtained from $\hat{B}$.

After the feature of size $H/4 \times W/4 \times 48$ is input to the first stage, the output size is $H/4 \times W/4 \times C$. Then, for each subsequent stage, the feature size is halved, and the number of channels is doubled. Therefore, the outputs of the four stages are $T_1 \in \mathbb{R}^{H/4 \times W/4 \times C}$, $T_2 \in \mathbb{R}^{H/8 \times W/8 \times 2C}$, $T_3 \in \mathbb{R}^{H/16 \times W/16 \times 4C}$, and $T_4 \in \mathbb{R}^{H/32 \times W/32 \times 8C}$.

We input the four features $T_1 \sim T_4$ into the global average pooling layer to obtain the feature vector $T_1' \sim T_4'$. The dimensions of the feature vectors are $1 \times 1 \times C$, $1 \times 1 \times 2C$, $1 \times 1 \times 4C$, and $1 \times 1 \times 8C$. By flattening all the feature vectors T_1' $\sim T_4'$, the dimensionality of the feature vectors becomes C, $2C$, $4C$, and $8C$. For the stability of training, all the feature vectors are normalized by the BatchNorm layer. All the feature vectors $T_1' \sim T_4'$ are spliced to obtain a global feature vector of dimension $15C$, which is fed into the softmax layer for emotion recognition.

C. Loss Function

Considering both local feature branches and global feature branches, the multibranch loss function L_{multi} contains two terms:

$$L_{multi} = (1-\sigma_1) \times L_{local} + \sigma_1 \times L_{global}, \quad (12)$$

where L_{local} denotes the loss function of the local feature branch and L_{global} denotes the loss function of the global feature branch. σ_1 was set to 0.8 in the experiment. Both loss functions use label-smoothing cross-entropy, and the calculation process is presented as follows:

$$Loss = -\sum_{i=1}^{K} y_i \log p_i, \quad (13)$$

$$y_i = \begin{cases} (1-\varepsilon) & if\,(i = y) \\ \dfrac{\varepsilon}{K-1} & if\,(i \neq y) \end{cases}, \quad (14)$$

$$p_i = \frac{\exp(z_i)}{\sum_{j=1}^{K} \exp(z_j)}, \quad (15)$$

where K denotes the total number of categories, ε is a small hyperparameter, p_i denotes the prediction result (obtained by performing softmax calculations on the output vector), y_i denotes the true label value, and $y_i = 0,1,2,\cdots,K$ denotes the true image category.

To ensure that the output distribution of the local feature branch is consistent with that of the global feature branch, a constraint term is added between the two branches in the loss function. The formula is presented as follows:

$$L_{dis} = (1-\sigma_2) \times L_{local\_to\_global} + \sigma_2 \times L_{global\_to\_local}, \quad (16)$$

where $L_{local\_to\_global}$ denotes the distribution loss to which the global feature branch output is targeted and the local feature branch output is fitted. $L_{global\_to\_local}$ denotes the distribution loss to which the global feature branch output is fitted, with the local feature branch output as the target. σ_1 was set to 0.8 in the experiment. The corresponding formulas are presented as follows:

$$L_{local\_to\_global} = \sum_{i=1}^{K} y_i \log\left(\frac{p_i^{local}}{p_i^{global}}\right), \quad (17)$$

$$L_{global\_to\_local} = \sum_{i=1}^{K} y_i \log\left(\frac{p_i^{global}}{p_i^{local}}\right), \quad (18)$$

where p_i^{local} denotes the probability of the ith class of the affective output from the local feature branch and p_i^{global} denotes the probability of the ith class of the affective output from the global feature branch.

To ensure the accuracy of the class activation map, the image class activation map loss, L_{CAM} is introduced to the loss function to guide the generation of the emotional class activation map. The calculation is presented as follows:

$$L_{CAM} = \left(-\sum_{i=1}^{K} y_i \log(p_i^{cam})\right) + \left(\sum_{i=1}^{K} y_i \log\left(\frac{p_i^{cam}}{p_i^{global}}\right)\right), \quad (19)$$

where p_i^{cam} denotes the probability of the *ith* class of the emotion output from the image emotion class activation map branch. The network is branched to obtain more accurate maps by learning global features to extract branch outputs and true labels.

The loss function of the entire network consists of L_{multi}, L_{dis} and L_{CAM} components, which are expressed as

$$L_{cls} = \lambda_1 \times L_{multi} + \lambda_2 \times L_{dis} + \lambda_3 \times L_{CAM} \quad (20)$$

The entire network is trained in an end-to-end manner. Local feature branches work in tandem with global feature branches. The use of image emotion activation map branches improves recognition performance with additional supervisory information. As the performance improves, the accuracy of the image emotion activation map also improves, which improves the recognition performance of the network.

VI. EXPERIMENTS

Using an image emotion recognition task, this study demonstrated that the proposed network is effective for image emotion recognition.

A. Experimental Setup

1) **Datasets**

The experiments were based on three datasets.

a) The Emotion-6 dataset [58] consists of 8,350 images culled from 150K images crawled by Google and Flickr. According to the basic human emotion theory proposed by Ekman, the images in this dataset were labeled as six emotions: anger, fear, joy, sadness, love, and surprise. As in [51], 80% of the images in the dataset were randomly selected for training, and the remainder were used for testing.

b) The FI-8 dataset [17] is a collection of images from Flickr and Instagram. In total, 23,308 images were separately labeled using up to eight emotional tags [51]. In total, 80%, 15%, and 5% of the images were randomly selected for training, testing, and validation.

c) The WEBEmo dataset [58] consists of 268K images collected from the web; it is the largest public dataset available for image emotion detection. According to Parrott's emotion hierarchy model, the dataset was divided into 25 emotion categories and 6 emotion categories, named WEBEmo-25 and WEBEmo-6, respectively. In total, 80% of the images were selected for training, and the remainder were selected for testing.

2) **Implementation Details**

Experiments were conducted in a PyTorch1.10 environment, and all parameters in the network were initialized using the model pretrained with ImageNet. λ_1, λ_2, and λ_3 were set to 0.8, 0.1, and 0.1, respectively. Stochastic gradient descent was employed for network training. The initial learning rate was set to 0.01, and its decay was determined using cosine annealing. The batch size was set to 16, and all experiments were conducted on four NVIDIA GTX TITAN V blocks with 12-GB video memory per GPU. All short edges of the training images were resized to 256, and the original aspect ratio was kept constant. Subsequently, the cropped images served as input to the network. Each channel of all input images was normalized so that both its mean and variance were zero.

B. Results

1) **Results for FI-8**

This study conducted two- and eight-class experiments with the FI-8 dataset. The models using Swin-small and Swin-base as the backbone networks are named CGLF-Net (small) and CGLF-Net (base), respectively.

In Table I, the results of the proposed algorithm are compared with those of the current common networks on the FI-8 dataset. As shown, the proposed CGLF-Net (base) algorithm worked best on the 2-class FI dataset, with an accuracy of 95.24%. Moreover, on the 8-class FI dataset, the proposed CGLF-Net (base) algorithm had an accuracy of 75.61%, which is substantially better than the traditional CNN.

Fig. 8 shows the confusion matrix of the proposed network structure for the FI-8 dataset. In the confusion matrix, the most confusing category was disgust. The proposed algorithm had the highest recognition rate of 84% for amusement. Only disgust had a recognition rate below 55% among the eight categories, and the recognition rate of the sad category was not high.

TABLE I
RECOGNITION ACCURACY OF THE PROPOSED NETWORK AND NETWORKS FROM THE LITERATURE WITH THE FI-8 DATASET

Model (Proper Name)	**#params**	**Recognition Accuracy (%)**	
		FI-2	**FI-8**
SentiBank [41]	-	56.47	44.49
Zhao et al. [12]	-	-	46.52
DeepSentiBank [44]	-	64.39	53.16
PCNN	-	75.34	56.16
AlexNet	61.10 M	72.43	58.30
ResNet-50	23.52 M	85.43	64.74
ResNet-101	42.52 M	85.92	64.53
ResNet-152	58.12 M	-	68.64
ViT-base [25]	86 M	89.65	69.74
Swin-base [26]	88 M	85.50	70.80
WSCNet [50]	-	-	70.07
Zhu et al. [54]	-	-	73.03
Rao et al. + L_{cls} [46]	48.80 M	-	73.05
Rao et al. + L_{multi} [46]	48.80 M	87.51	75.46
Zhang et al. + ResNet-50 (1-crop) [51]	31.29 M	89.83	73.80
Zhang et al. + ResNet-50 (10-crop) [51]	31.29 M	90.58	74.80
Zhang et al. + ResNet-101 (1-crop) [51]	50.29 M	90.35	75.05
Zhang et al. + ResNet-101 (10-crop) [51]	50.29 M	90.97	75.91
CGLF-Net (small)	59.53 M	94.48	74.13
CGLF-Net (Base)	96.68 M	95.24	75.61

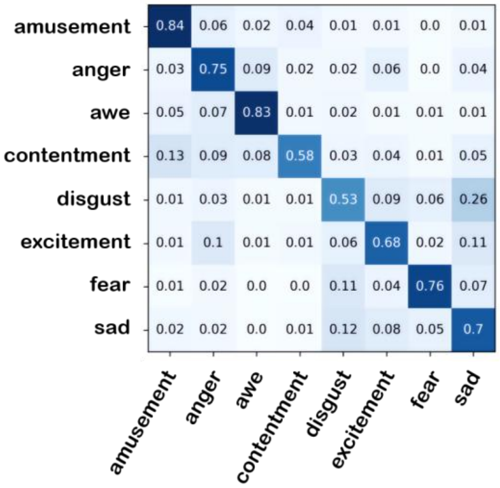


Fig. 8. Confusion matrix of the performance of the proposed CGLF-Net (base) with the FI-8 dataset.

Rao et al. [46] proposed a multilevel region-based CNN framework for image emotion classification recognition. Their proposed method uses loss to train the ResNet101 network, achieving a recognition rate of 75.46% for the 8-class FI dataset and 87.51% for the 2-class FI dataset. The proposed CGLF-Net (Base) approach improved performance by 2.56 and 7.73%. Moreover, the model proposed in [46] must first be trained on a dataset containing region annotations to detect emotionally

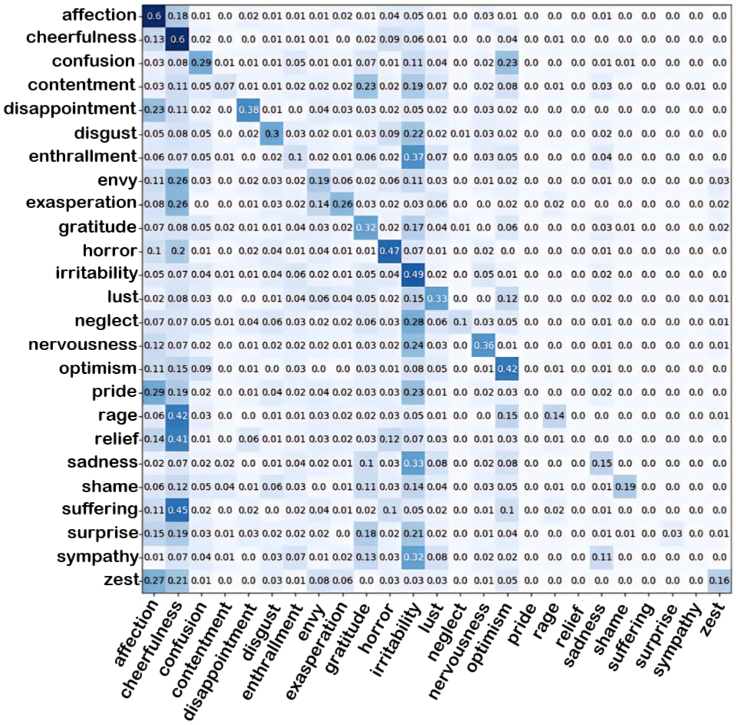


Fig. 9. Confusion matrix of the performance of the proposed CGLF-Net (base) with the WEBEmo-25 dataset.

TABLE II
RECOGNITION ACCURACY OF THE PROPOSED NETWORK WITH THE WEBEMO-6 AND WEBEMO-25 DATASETS

Model (Proper Name)	#params	Recognition Accuracy (%)	
		WEBEmo-6	WEBEmo-25
ResNet-50	23.52 M	50.38	30.61
ResNet-101	42.52 M	50.71	30.95
ResNet-152	58.12 M	51.15	31.07
ViT-base [25]	86 M	53.68	31.43
Swin-base	88 M	55.32	34.28
Zhang et al. + ResNet-50 (1-crop) [51]	31.29 M	51.90	32.53
Zhang et al. + ResNet-50 (10-crop) [51]	31.29 M	53.06	32.54
Zhang et al. + ResNet-101 (1-crop) [51]	50.29 M	52.86	30.05
Zhang et al. + ResNet-101 (10-crop) [51]	50.29 M	53.88	33.01
CGLF-Net (small)	59.53 M	56.13	36.30
CGLF-Net (Base)	96.68 M	**56.70**	**36.61**

TABLE III
RECOGNITION ACCURACY OF THE PROPOSED NETWORK WITH THE EMOTION-6 DATASET.

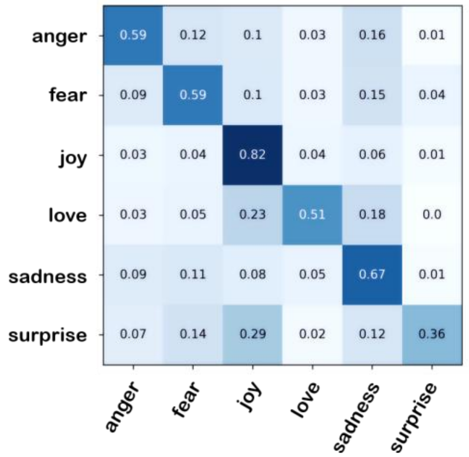
Model (Proper Name)	#params	Recognition Accuracy (%)
ResNet-50	23.51 M	53.97
ResNet-101	42.52 M	56.16
ResNet-152	58.16 M	58.72
ViT-Base [25]	86 M	57.74
Swin-small	50 M	55.04
Swin-based	88 M	54.68
Zhang et al. + ResNet-50 (1-crop) [51]	31.29 M	58.86
CGLF-Net (small)	59.53 M	63.27
CGLF-Net (Base)	96.68 M	**65.01**

significant regions. However, most current datasets are not labeled with emotionally significant regions. The network in [51] achieved accuracies of 75.05 and 90.35% for the 8-class FI dataset and 2-class FI dataset, respectively, when training the ResNet101 model using the 1-crop approach. By comparison, the proposed network improved accuracy by 0.56 and 4.89% for the 8-class FI dataset and 2-class FI dataset, respectively, when the same 1-crop approach was employed for model training. All features of the model in [51] were directly obtained from the ResNet network, which caused the model to excessively focus on local regions and disregard global feature information, owing to the focus of CNNs on local features during extraction. In this study, the model uses a Swin Transformer to extract the global features of the image, and the transformer structure enables it to obtain a more robust image feature representation.

As shown in Table IV, we additionally compare the precision, recall and F1 scores of different algorithms on the FI-8 dataset. Compared with the traditional methods, our proposed algorithm is substantially ahead in all three metrics. Compared with the work of Zhang et al. [51], the precision of our designed algorithm is slightly lower because the 10-crop technique utilized by Zhang et al. [51] expands the dataset by a factor of

TABLE IV
PRECISION, RECALL AND F1 SCORES OF THE PROPOSED NETWORK AND NETWORKS FROM THE LITERATURE WITH THE EMOTION-6, FI-8 AND WEBEMO-25 DATASET

Model	#params	Emotion-6			FI-8			WEBEmo-25		
		precision	recall	F1	precision	recall	F1	precision	recall	F1
Resnet-18	11.7 M	0.356	0.350	0.353	0.395	0.410	0.402	-	-	-
Resnet-101	42.52 M	0.373	0.360	0.366	0.466	0.431	0.448	-	-	-
Inception-v4 [59]	163 M	0.368	0.352	0.360	0.479	0.417	0.446			
Yang et al. [60]	-	0.327	0.327	0.327	0.450	0.428	0.439			
AlexNet	61.10 M	0.302	0.304	0.303	0.5201	0.5050	0.5124	-	-	-
MldrNet [39]	-	-	-	-	0.5964	0.5829	0.5896	-	-	-
MG [61]	-	-	-	-	0.5738	0.5749	0.5842			
DMN-HS [62]	-	0.447	0.454	0.450	0.497	0.556	0.525	-	-	-
Zhu et al. [54]	-	-	-	-	0.6514	0.6495	0.6504	-	-	-
Rao et al.[46]	48.80 M	-	-	-	0.6952	0.6788	0.6856	-	-	-
Zhang et al. [51]	50.29 M	0.6041	0.5639	0.5833	**0.7208**	0.6939	0.7071	0.0781	0.0843	0.0811
CGLF-Net (Base)	96.68 M	**0.6351**	**0.5879**	**0.6106**	0.7198	**0.7082**	**0.7140**	**0.238**	**0.3227**	**0.2740**



	anger	fear	joy	love	sadness	surprise
anger	0.59	0.12	0.1	0.03	0.16	0.01
fear	0.09	0.59	0.1	0.03	0.15	0.04
joy	0.03	0.04	0.82	0.04	0.06	0.01
love	0.03	0.05	0.23	0.51	0.18	0.0
sadness	0.09	0.11	0.08	0.05	0.67	0.01
surprise	0.07	0.14	0.29	0.02	0.12	0.36

Fig. 10. Confusion matrix of the performance of the proposed CGLF-Net (base) with the Emotion-6 dataset.

TABLE V
IMPACT OF DIFFERENT CONFIGURATIONS OF THE LOCAL FEATURE EXTRACTION BRANCH ON RECOGNITION ACCURACY: RESNET-50 BASELINE

Method	Recognition Accuracy (%)	
	FI-8	Emotion-6
Baseline	64.74	53.97
Baseline + Multiscale	65.65	55.64
Baseline + Multiscale + CAM	66.34	56.96
Baseline + Multiscale + CAM + Local attention	**68.00**	**58.22**

Note: CAM = Class activation map.

TABLE VI
IMPACT OF DIFFERENT NETWORK STRUCTURES ON THE RECOGNITION ACCURACY WITH THE EMOTION-6 AND FI-8 DATASETS

Method	Recognition Accuracy (%)	
	FI-8	Emotion-6
ResNet-50	64.74	53.97
Only global feature branch (Swin-based)	70.80	54.68
ResNet-50 + Global feature branch (Swin-based)	72.76	60.74
Our local feature branch + Global feature branch (Swin-based)	**75.61**	**65.01**

10, but the recall and F1 score are still improved by 1.43% and 0.69%, respectively.

2) **Results for WEBEmo**

On the WEBEmo-6 and WEBEmo-25 datasets, the proposed algorithm was compared to contemporary algorithms. The results are presented in Table II. Compared with ResNet-50, the proposed CGLF-Net (base) algorithm improved the recognition rate by 6.32 and 6.00% with the respective datasets when using ResNet-50 as a local feature branch. Without the 10-crop approach, the accuracy of the algorithm proposed in [51] was 52.86 and 30.05% for the WEBEmo-6 dataset and WEBEmo-25 dataset, respectively. The proposed algorithm improved its accuracy by 3.84 and 6.56%. Furthermore, the proposed CGLF-Net (Base) algorithms improved the recognition rates by 2.82 and 3.60%, compared with the work of Zhang et al. [51] using the 10-crop approach.

Fig. 9 shows the confusion matrix of the proposed CGLF-Net (base) algorithms using the WEBEmo25 dataset. These categories were most easily confused with the irritability category. The emotions with the highest recognition rates were cheerfulness and affection (60%). Among the 25 categories, six had recognition rates below 10%. Of these, four categories had zero recognition: pride, relief, suffering, and sympathy.

The proposed method is better than the work of Zhang et al. [51] in terms of precision, recall, and F1 scores across the board, and the results are shown in Table IV. This finding may be attributed to the stronger performance of the transformer network compared to the CNN network on large-scale datasets.

3) **Results on Emotion-6**

The experimental results on the Emotion-6 dataset are listed in Table III. Compared with the conventional ResNet-50 and ResNet-101 networks, the proposed CGLF-Net (small) improved accuracy by 9.3 and 7.11%, respectively. However, the proposed CGLF-Net (base) architectures improved

TABLE VII
IMPACT OF INTENSITY LOSS FUNCTION ON RECOGNITION ACCURACY WITH THE EMOTION-6 AND FI-8 DATASETS; RESNET-50 AND SWIN-BASED BACKBONES

Loss function	Recognition Accuracy (%)	
	FI-8	Emotion-6
L_{multi}	73.32	61.36
$\lambda_1 \times L_{multi} + \lambda_2 \times L_{dis}$	75.08	64.08
$\lambda_1 \times L_{multi} + \lambda_3 \times L_{CAM}$	74.04	61.96
$\lambda_1 \times L_{multi} + \lambda_2 \times L_{dis} + \lambda_3 \times L_{CAM}$	**75.61**	**65.01**

accuracy by 11.04, 8.85, and 6.29% compared to the conventional ResNet-50, ResNet-101, and ResNet-152 networks, respectively. Moreover, the results of the proposed algorithms were all better than those of a single transformer network. In particular, the proposed CGLF-Net (base) achieved an accuracy improvement of 10.33% compared to the Swin-base algorithm.

The confusion matrix based on the CGLF-Net (base) algorithms is shown in Fig. 10. The surprise category was the most confusing of the six categories, with a recognition rate of 36%. The recognition rates for the remaining categories exceeded 50%, with the highest rate for the joy category (82%).

We compared the precision, recall, and F1 scores of the current common methods on the Emotion-6 dataset. As shown by the results in Table IV, our method achieves the highest precision, recall, and F1 scores, with 3.1%, 2.4%, and 2.73% improvements in the corresponding values compared to the work of Zhang et al. [51] This finding shows that our algorithm performs equally well on a small dataset.

4) Ablation Experiment

To validate the efficacy of the proposed method, an ablation analysis was performed on two datasets: FI-8 and Emotion-6. In the experiments, the effects of different network structures and the proposed loss functions on the recognition results were separately investigated. The results are shown in Tables V, VI, and VII.

First, an ablation analysis was performed on the multiscale, CAM, and LAM modules in the local feature branch of the network model. ResNet-50 was selected as the baseline network, and its performance on the FI-8 and Emotion-6 datasets was considered the baseline. The results are presented in Table V. Notably, the architecture with all modules achieved the best performance.

The idea of the emotion CAM module was inspired by the work of [63], which indicated that the module could automatically localize the overall object of semantic interest. The LAM computes attention from spatial and channel aspects, and the module enables the network to focus on the emotion information in the image. The experimental results demonstrated that the individual modules employed in the local feature branch are effective for the final image emotion recognition.

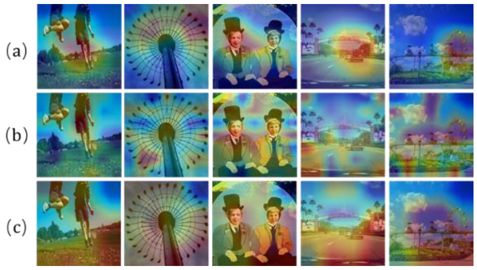


Fig. 11. Attention graph samples generated by different networks: (a) ResNet-50, (b) swin-based network, and (c) proposed model.

Moreover, this study analyzed the performance of different network structures on the FI-8 and Emotion-6 datasets, the results of which are presented in Table VI. As shown, the results were not satisfactory when only ResNet-50 was used for image sentiment recognition on both datasets. On the FI-8 dataset, the Swin-based network improved accuracy by 6.06% compared with the ResNet-50 network. However, on the Emotion-6 dataset, the accuracy improvement was only 0.71% because the Swin-based network with a transformer structure is more sensitive to the sample size and performs worse on smaller datasets. Furthermore, the results show that the recognition accuracy of the structure combining the ResNet-50 and Swin-based networks was higher than that of the structure using only a single network. This finding demonstrates that the method of combining the local features extracted by ResNet-50 with the global features extracted by the Swin-base model is effective.

Image sentiment recognition was improved after replacing the ResNet-50 network with the proposed local feature extraction branch (Table VI). This finding indicates that the local feature extraction branch designed in this study can be used to more effectively extract the local features of images than ResNet-50.

To determine the impact of the loss function on recognition accuracy, ResNet-50 and Swin-base models were selected as the baseline, following the network structure proposed in this study. The results are presented in Table VII.

The main reason for the poor recognition performance when L_{multi} serves as the loss function is that only the respective losses of the local and global feature branches are considered, and the relationship between the two branches is disregarded. The addition of L_{dis} to the loss function significantly improved recognition performance, which demonstrates that the relationship between two different branches should be considered in a two-branch network. Notably, L_{CAM} as additional supervision information further improved recognition accuracy. These results demonstrate that the proposed loss function effectively improves algorithm performance.

5) Visualization

Fig. 11 presents a few samples of the attention maps generated with different network structures. Traditional CNNs clearly focus more on local regions when recognizing emotions; thus, their network attention is mainly focused on certain local regions. In contrast, the attention of transformer networks is more decentralized. By comparison, the attention of the proposed model covers most regions of an image as well as the prominent objects and the image background. This finding shows that in our proposed network, local and global regions significantly contribute to emotion recognition. This finding indicates that the network meets our original design intention of combining local and global features of images for emotion recognition.

V. Conclusion

In this paper, we proposed a novel CGLF-Net model that combines global self-attention and local multiscale feature extraction based on CNNs and transformers to explore the integration of global and local features for achieving superior image emotion recognition. The CGLF-Net network consists of two branches: a local emotional feature branch and a global emotional feature branch. In the local emotional feature branch, multiscale features of an image are obtained using a modified feature pyramid, and local features are enhanced by CAMs that weight the local attention module based on the composition of CAMs and CBAMs, which makes the attention of the network focus on local image details. In the global emotional feature branch, the global feature representation of the image is extracted by replacing the convolution operation with a self-attention mechanism, where the global feature representation of the model is enhanced by using a cross-scale feature fusion module. Eventually, the network was trained using the multibranch loss function proposed in this paper, which facilitates the combination of local and global emotional feature branches to obtain a comprehensive image emotion feature representation. The results of ablation experiments show that, compared with the features extracted by traditional CNNs, the local multiscale features effectively improve the emotion recognition rate of CGLF-Net. Moreover, the results demonstrate that the model that uses a combination of global and local features outperforms the model using only one type of feature. Compared with existing state-of-the-art algorithms, the CGLF-Net framework designed in this paper achieves promising results on multiple benchmark datasets. In particular, the performance on the Emotion-6 dataset reaches the highest level.

However, the proposed algorithm needs to compute all regions of the image based on the self-attention mechanism in the global feature branch, which results in a significant increase in the number of network parameters compared to the traditional CNN model. In future works on image emotion recognition, attempts will be made to more efficiently fuse the various features.

REFERENCES

[1] P.J. Lang, "*A bio-informational theory of emotional imagery*". Psychophysiology, 1979. **16**(6): p. 495-512

[2] B.H. Detenber, R.F. Simons, and G.G. Bennett Jr, "*Roll 'em!: The effects of picture motion on emotional responses*". Journal of Broadcasting & Electronic Media, 1998. **42**(1): p. 113-127

[3] P.J. Lang, M.M. Bradley, and B.N. Cuthbert, "*Emotion, motivation, and anxiety: Brain mechanisms and psychophysiology*". Biological psychiatry, 1998. **44**(12): p. 1248-1263

[4] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J.Z. Wang, J. Li, and J. Luo, "*Aesthetics and emotions in images*". IEEE Signal Processing Magazine, 2011. **28**(5): p. 94-115

[5] L. Pang, S. Zhu, and C.-W. Ngo, "*Deep multimodal learning for affective analysis and retrieval*". IEEE Transactions on Multimedia, 2015. **17**(11): p. 2008-2020

[6] L. Tong, R. Tong, and L. Chen, "*Efficient retrieval algorithm for multimedia image information*". Multimedia Tools and Applications, 2020. **79**: p. 9469-9487

[7] S. Qian, T. Zhang, and C. Xu. "*Multi-modal multi-view topic-opinion mining for social event analysis*". in *Proceedings of the 24th ACM international conference on Multimedia*. 2016.

[8] Q. You, J. Luo, H. Jin, and J. Yang. "*Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia*". in *Proceedings of the Ninth ACM international conference on Web search and data mining*. 2016.

[9] S. Zhao, H. Yao, Y. Gao, R. Ji, W. Xie, X. Jiang, and T.-S. Chua. "*Predicting personalized emotion perceptions of social images*". in *Proceedings of the 24th ACM international conference on Multimedia*. 2016.

[10] B. Wu, J. Jia, Y. Yang, P. Zhao, J. Tang, and Q. Tian, "*Inferring emotional tags from social images with user demographics*". IEEE Transactions on Multimedia, 2017. **19**(7): p. 1670-1684

[11] X. Lu, Z. Lin, H. Jin, J. Yang, and J.Z. Wang. "*Rapid: Rating pictorial aesthetics using deep learning*". in *Proceedings of the 22nd ACM international conference on Multimedia*. 2014.

[12] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun. "*Exploring principles-of-art features for image emotion recognition*". in *Proceedings of the 22nd ACM international conference on Multimedia*. 2014. [DOI: 10.1145/2647868.2654930].

[13] K.-C. Peng, T. Chen, A. Sadovnik, and A.C. Gallagher. "*A mixed bag of emotions: Model, predict, and transfer emotion distributions*". in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[14] J. Yang, M. Sun, and X. Sun. "*Learning visual sentiment distributions via augmented conditional probability neural network*". in *Proceedings of the AAAI Conference on Artificial Intelligence*. 2017.

[15] H.-B. Kang. "*Affective content detection using HMMs*". in *Proceedings of the eleventh ACM international conference on Multimedia*. 2003.

[16] X. Yao, D. She, H. Zhang, J. Yang, M.-M. Cheng, and L. Wang, "*Adaptive deep metric learning for affective image retrieval and classification*". IEEE Transactions on Multimedia, 2020. **23**: p. 1640-1653

[17] Q. You, J. Luo, H. Jin, and J. Yang. "*Building a large scale dataset for image emotion recognition: The fine print and the benchmark*". in *Proceedings of the AAAI conference on artificial intelligence*. 2016.

[18] R.J. Compton, "*The interface between emotion and attention: A review of evidence from psychology and neuroscience*". Behavioral and cognitive neuroscience reviews, 2003. **2**(2): p. 115-129

[19] S. Zhao, X. Yao, J. Yang, G. Jia, G. Ding, T.-S. Chua, B.W. Schuller, and K. Keutzer, "*Affective image content analysis: Two decades review and new perspectives*". IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021 [DOI: 10.1109/TPAMI.2021.3094362].

[20] B. Li, W. Xiong, W. Hu, and X. Ding. "*Context-aware affective images classification based on bilayer sparse representation*". in *Proceedings of the 20th ACM international conference on Multimedia*. 2012.

[21] M. Chen, L. Zhang, and J.P. Allebach. "*Learning deep features for image emotion classification*". in *2015 IEEE International Conference on Image Processing (ICIP)*. 2015. IEEE

[22] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye. "*Conformer: Local features coupling global representations for visual recognition*". in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

[23] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang. "*Cvt: Introducing convolutions to vision transformers*". in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

[24] Y. Chen, X. Dai, D. Chen, M. Liu, X. Dong, L. Yuan, and Z. Liu. "*Mobile-former: Bridging mobilenet and transformer*". in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin, "*Attention is all you need*". Advances in neural information processing systems, 2017. **30**

[26] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. "*Swin transformer: Hierarchical vision transformer using shifted windows*". in *Proceedings of the IEEE/CVF International Conference on Computer*

Vision. 2021.

[27] L. Camras, *Emotion: a psychoevolutionary synthesis*. 1980, JSTOR.

[28] P. Ekman, *"An argument for basic emotions"*. Cognition & emotion, 1992. **6**(3-4): p. 169-200 [DOI: 10.1080/02699939208411068].

[29] W.G. Parrott, *"Emotions in social psychology: Essential readings"*. 2001: psychology press.

[30] J.A. Mikels, B.L. Fredrickson, G.R. Larkin, C.M. Lindberg, S.J. Maglio, and P.A. Reuter-Lorenz, *"Emotional category data on images from the International Affective Picture System"*. Behavior research methods, 2005. **37**(4): p. 626-630 [DOI: 10.3758/bf03192732].

[31] H. Schlosberg, *"Three dimensions of emotion"*. Psychological review, 1954. **61**(2): p. 81

[32] J. Lee and E. Park, *"Fuzzy similarity-based emotional classification of color images"*. IEEE Transactions on Multimedia, 2011. **13**(5): p. 1031-1039

[33] S. Zhao, H. Yao, Y. Gao, R. Ji, and G. Ding, *"Continuous probability distribution prediction of image emotions via multitask shared sparse regression"*. IEEE transactions on multimedia, 2016. **19**(3): p. 632-645

[34] J. Machajdik and A. Hanbury. *"Affective image classification using features inspired by psychology and art theory"*. in *Proceedings of the 18th International Conference on Multimedea 2010, Firenze, Italy, October 25-29, 2010*. 2010. [DOI: 10.1145/1873951.1873965].

[35] W. Wei-Ning, Y. Ying-Lin, and J. Sheng-Ming. *"Image retrieval by emotional semantics: A study of emotional space and feature extraction"*. in *2006 IEEE International Conference on Systems, Man and Cybernetics*. 2006. IEEE [DOI: 10.1109/ICSMC.2006.384667].

[36] V. Yanulevskaya, J.C. van Gemert, K. Roth, A.-K. Herbold, N. Sebe, and J.-M. Geusebroek. *"Emotional valence categorization using holistic image features"*. in *2008 15th IEEE international conference on Image Processing*. 2008. IEEE

[37] J. Yuan, S. Mcdonough, Q. You, and J. Luo. *"Sentribute: image sentiment analysis from a mid-level perspective"*. in *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*. 2013. [DOI: 10.1145/2502069.2502079].

[38] X. Wang, J. Jia, J. Yin, and L. Cai. *"Interpretable aesthetic features for affective image classification"*. in *2013 IEEE International Conference on Image Processing*. 2013. IEEE

[39] T. Rao, X. Li, and M. Xu, *"Learning multi-level deep representations for image emotion classification"*. Neural processing letters, 2020. **51**(3): p. 2043-2061 [DOI: 10.1007/s11063-019-10033-9].

[40] X. Lu, R.B. Adams, J. Li, M.G. Newman, and J.Z. Wang. *"An investigation into three visual characteristics of complex scenes that evoke human emotion"*. in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2017. IEEE

[41] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. *"Large-scale visual sentiment ontology and detectors using adjective noun pairs"*. in *Proceedings of the 21st ACM international conference on Multimedia*. 2013. [DOI: 10.1145/2502081.2502282].

[42] B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara, and S.-F. Chang. *"Visual affect around the world: A large-scale multilingual visual sentiment ontology"*. in *Proceedings of the 23rd ACM international conference on Multimedia*. 2015. [DOI: 10.1145/2733373.2806246].

[43] A.R. Ali, U. Shahid, M. Ali, and J. Ho. *"High-level concepts for affective understanding of images"*. in *2017 IEEE winter conference on applications of computer vision (WACV)*. 2017. IEEE [DOI: 10.1109/WACV.2017.81].

[44] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, *"Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks"*. arXiv preprint arXiv:1410.8586, 2014

[45] Q. You, J. Luo, H. Jin, and J. Yang. *"Robust image sentiment analysis using progressively trained and domain transferred deep networks"*. in *Twenty-ninth AAAI conference on artificial intelligence*. 2015. [DOI: 10.1609/aaai.v29i1.9179].

[46] T. Rao, X. Li, H. Zhang, and M. Xu, *"Multi-level region-based convolutional neural network for image emotion classification"*. Neurocomputing, 2019. **333**: p. 429-439 [DOI: 10.1016/j.neucom.2018.12.053].

[47] Q. You, H. Jin, and J. Luo. *"Visual sentiment analysis by attending on local image regions"*. in *Proceedings of the AAAI conference on artificial intelligence*. 2017.

[48] 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, 2018. **20**(9): p. 2513-2525 [DOI: 10.1109/TMM.2018.2803520].

[49] S. Zhao, Z. Jia, H. Chen, L. Li, G. Ding, and K. Keutzer. *"Pdanet: Polarity-consistent deep attention network for fine-grained visual emotion regression"*. in *Proceedings of the 27th ACM international conference on multimedia*. 2019.

[50] 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, 2019. **22**(5): p. 1358-1371 [DOI: 10.1109/TMM.2019.2939744].

[51] H. Zhang and M. Xu, *"Weakly supervised emotion intensity prediction for recognition of emotions in images"*. IEEE Transactions on Multimedia, 2020. **23**: p. 2033-2044 [DOI: 10.1109/TMM.2020.3007352].

[52] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. *"Training data-efficient image transformers & distillation through attention"*. in *International conference on machine learning*. 2021. PMLR

[53] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. *"Pyramid vision transformer: A versatile backbone for dense prediction without convolutions"*. in *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

[54] X. Zhu, L. Li, W. Zhang, T. Rao, M. Xu, Q. Huang, and D. Xu. *"Dependency Exploitation: A Unified CNN-RNN Approach for Visual Emotion Recognition"*. in *IJCAI*. 2017.

[55] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. *"Feature pyramid networks for object detection"*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[56] J. Yang, D. She, Y.-K. Lai, and M.-H. Yang. *"Retrieving and classifying affective images via deep metric learning"*. in *Proceedings of the AAAI Conference on Artificial Intelligence*. 2018. [DOI: 10.1609/aaai.v32i1.11275].

[57] S. Woo, J. Park, J.-Y. Lee, and I.S. Kweon. *"Cbam: Convolutional block attention module"*. in *Proceedings of the European conference on computer vision (ECCV)*. 2018.

[58] R. Panda, J. Zhang, H. Li, J.-Y. Lee, X. Lu, and A.K. Roy-Chowdhury. *"Contemplating visual emotions: Understanding and overcoming dataset bias"*. in *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

[59] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. *"Inception-v4, inception-resnet and the impact of residual connections on learning"*. in *Proceedings of the AAAI conference on artificial intelligence*. 2017. [DOI: 10.1609/aaai.v31i1.11231].

[60] J. Yang, D. She, and M. Sun. *"Joint Image Emotion Classification and Distribution Learning via Deep Convolutional Neural Network"*. in *IJCAI*. 2017.

[61] L. Lim, H.-Q. Khor, P. Chaemchoy, J. See, and L.-K. Wong. *"Where is the emotion? Dissecting a multi-gap network for image emotion classification"*. in *2020 IEEE International Conference on Image Processing (ICIP)*. 2020. IEEE [DOI: 10.1109/ICIP40778.2020.9191258].

[62] Y. Liang, K. Maeda, T. Ogawa, and M. Haseyama. *"Deep metric network via heterogeneous semantics for image sentiment analysis"*. in *2021 IEEE International Conference on Image Processing (ICIP)*. 2021. IEEE [DOI: 10.1109/ICIP42928.2021.9506701].

[63] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T.S. Huang. *"Adversarial complementary learning for weakly supervised object localization"*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Yutong Luo received a B.S. degree in software security from Hangzhou Dianzi University, Hangzhou, China, in 2016, and a master's degree in software engineering from Chongqing Normal University, Chongqing, China, in 2019. He is currently working toward a Ph.D. degree in computer science from Southwest University, Chongqing, China. His current research interests include affective computing, computer vision, and machine learning.

Xinyue Zhong received a bachelor's degree (with honors) in electrical engineering from the University of Tasmania and Southwest University in 2020. She is currently enrolled in an M.S. degree program in Information and Communication Engineering at Southwest University. Her research fields are affective computing, artificial intelligence, deep learning and EEG-based emotion recognition.

Minchen Zeng received a B.S. degree in electronic information engineering from Wuhan Polytechnic University, Wuhan, China, in 2017. He is currently working toward a master's degree in electronic information from Southwest University, Chongqing, China. His current research interests include affective computing, computer vision, and machine learning.

Jialan Xie received a B.S. degree in electronic information science and technology from Guizhou Education University, Guizhou, China, in 2015, and a Master's degree in signal and information processing from Southwest University, Chongqing, China, in 2018. She is currently working toward a Ph.D. degree in computer science from Southwest University, Chongqing, China. Her current research interests include affective computing, virtual reality and machine learning.

Shiyuan Wang received a B.S. degree in electronic information engineering from Minjiang University, Fuzhou, China, in 2016. She is currently working toward a Master's degree in signal and information processing from Southwest University, Chongqing, China. Her current research interests include affective computing, brain-computer interface, and microexpression recognition.

Guangyuan Liu, Jr. obtained a B.E. degree in physics from Southwest University, Chongqing, China, in 1983, and M.A. and Ph.D. degrees in electronic engineering from the University of Electronic Science and Technology in 1995 and 1999, respectively. He is currently a professor with the Chongqing Key Laboratory of Nonlinear Circuits and Intelligent Information Processing and the Assistant President of Southwest University. His research interests include affective computing, neural networks, computational intelligence, and fuzzy systems and applications.