Hindawi
Security and Communication Networks
Volume 2021, Article ID 6854586, 12 pages
https://doi.org/10.1155/2021/6854586

Research Article

A Saliency Detection and Gram Matrix Transform-Based Convolutional Neural Network for Image Emotion Classification

Zelin Deng,[1] **Qiran Zhu**,[1,2] **Pei He**,[3] **Dengyong Zhang**,[1] **and Yuansheng Luo**[1]

[1]*School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China*
[2]*School of Big Data and Artificial Intelligence, Xinyang University, Xinyang 464000, China*
[3]*School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou 510006, China*

Correspondence should be addressed to Dengyong Zhang; zhdy@csust.edu.cn

Received 28 May 2021; Accepted 2 August 2021; Published 10 August 2021

Academic Editor: Beijing Chen

Using the convolutional neural network (CNN) method for image emotion recognition is a research hotspot of deep learning. Previous studies tend to use visual features obtained from a global perspective and ignore the role of local visual features in emotional arousal. Moreover, the CNN shallow feature maps contain image content information; such maps obtained from shallow layers directly to describe low-level visual features may lead to redundancy. In order to enhance image emotion recognition performance, an improved CNN is proposed in this work. Firstly, the saliency detection algorithm is used to locate the emotional region of the image, which is served as the supplementary information to conduct emotion recognition better. Secondly, the Gram matrix transform is performed on the CNN shallow feature maps to decrease the redundancy of image content information. Finally, a new loss function is designed by using hard labels and probability labels of image emotion category to reduce the influence of image emotion subjectivity. Extensive experiments have been conducted on benchmark datasets, including FI (Flickr and Instagram), IAPSsubset, ArtPhoto, and Abstract. The experimental results show that compared with the existing approaches, our method has a good application prospect.

1. Introduction

Image sentiment analysis is becoming a research hotspot in the field of computer vision [1–6]. It is more difficult to analyze images at the emotional level compared with the recognition of objects in images [7–13] mainly because of the complexity and subjectivity of emotions [4]. First of all, due to the complexity of emotion, image emotion recognition work is to analyze the image at the emotional level, and the expression of emotion is also affected by numerous feature information [14], so it is difficult to design a discriminative representation feature to cover enough feature information, such as color, texture, and semantic information. Secondly, due to the subjectivity of image emotion, people with different lives and cultural backgrounds may have different emotional responses to the same image which makes it difficult to collect hard emotion labels of the image and lead to the uncertainty of the image's category label.

In previous studies, many researchers have proposed methods to solve the complexity and subjectivity of image emotion. For instance, Borth et al. [14] developed a visual sentiment ontology, which consisted of 1200 concepts and associated classifiers, and each concept was composed of an adjective expressing emotion and a noun related to object or scene. In the work of image emotion analysis, manual features, including color, texture, composition, balance, and harmony [2, 15, 16], are first used to analyze the emotion of the image. However, handmade features are unable to fully express the relationship between visual information and emotional arousal because handmade features cannot cover the important features related to image emotion [17].

Recently, researchers began using CNNs to solve difficult problems in image sentiment classification to further improve classification performance [1]. Different from the manual features, CNN can learn image representation in an end-to-end manner. Research results have proved that deep

CNN features are better than manual features in image emotion recognition [17]. However, due to the complexity and subjectivity of emotions, analyzing images at the emotional level is a more challenging task compared with traditional visual tasks, such as object classification and detection in the image. For the complexity of image emotion, most images can cause different emotional reactions, rather than a unique emotion. Previous studies mainly used visual features extracted from the global view of the image for emotion recognition, while ignored the fact that expression of image emotion mainly depends on the local regions of an image. Figure 1 shows the image samples and the main regions in them to evoke emotion. Obviously, some local regions of the image contain more emotional information than others. Besides, Alameda-Pineda et al. [18] pointed out that CNNs were unable to effectively extract emotional information from abstract paintings, which means emotions not only are induced by image semantics but also are conveyed through low-level visual features, such as texture, color, and shape.

In order to understand how CNNs designed for object recognition task works in image emotion recognition task, many studies on deep feature representation on convolutional neural network processing level have been conducted. Research shows that emotion recognition of the deep model is mainly based on semantic features of images, which can explain the successful application of CNN in image emotion recognition [2]. On the other hand, when the image is processed by the deeper CNN layers, the low-level visual features are gradually reduced. In some cases, people pay more attention to the background of the image than to the object in the image, that is, nonobject components may be more emotional than image contents [18]. This requires us to introduce the low-level visual features of the image when designing the classification features, but if we directly use the feature map obtained from the shallow network to describe the low-level visual features, there will be a problem of redundancy because the feature map also contains the image content information. Inspired by the work of image style transformation [19–21], we apply Gram matrix transformation on the feature maps from the shallow layers of the network to reduce the redundancy of image content.

In order to enhance the image emotion recognition performance, the CNN is proposed to improve with the following. Firstly, use the saliency detection method to extract the features of the local emotional regions to better invoke the emotions. Secondly, introduce multiple side branch structures in network to obtain the feature maps of the shallow layers and use the Gram matrix to transform the feature maps to decrease redundancy. Finally, design a new loss function by using the hard labels and probability labels of image emotion categories to reduce the impact of image emotion subjectivity on classification.

In summary, the contributions of our paper are summarized as follows:

(1) Use saliency detection algorithm to locate the emotional region in the image and extract the features of the emotional region in the image, which can avoid the noise information in the nonemotional region and give more attention to the local emotional regions.

(2) Design a method to calculate the Gram matrix of the feature map. After Gram matrix transformation, the redundancy of the image content information in the feature map is reduced, and new low-level visual features are obtained.

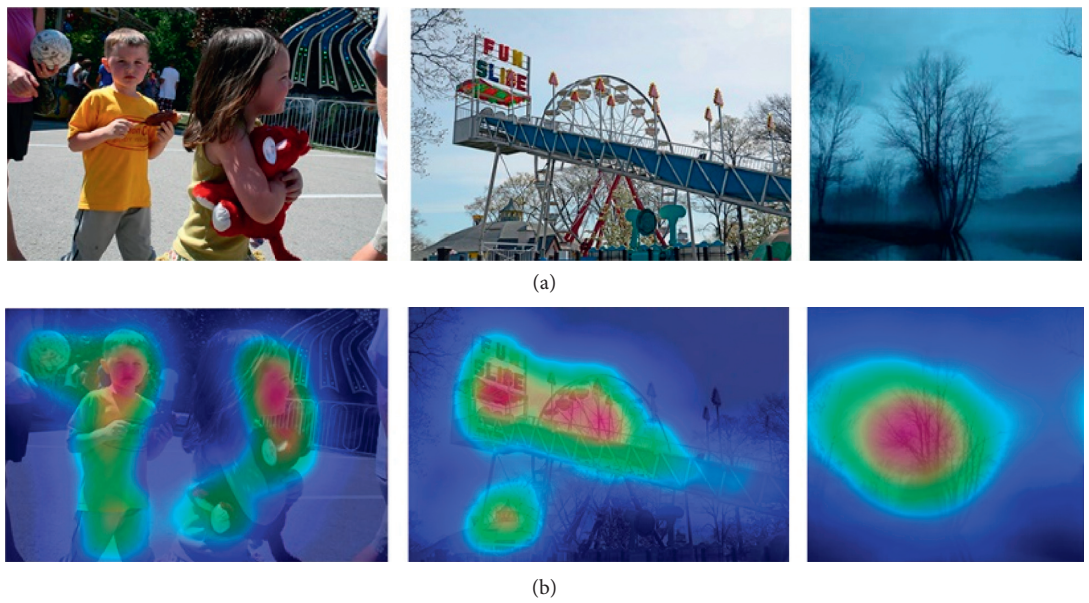
(3) Propose a new loss function by using the hard labels and probability labels of image emotion categories to reduce the impact of image emotion subjectivity on classification.

The remainder of this paper is as follows. In Section 2, we summarized and reviewed the related work of image emotion recognition and image saliency detection. Section 3 introduced our model and improvement work. Section 4 introduced the datasets used in the experiment and presented the experimental results and analysis of this work. In Section 5, our main work and future research keys are summarized.

2. Related Works

The analysis of images and videos on the emotional level has attracted the attention of more and more researchers [22–25], and a lot of research works have been carried out. In this section, we focus on reviewing the related work of image emotion analysis and image saliency detection.

2.1. Image Emotion Analysis. In the work of image sentiment classification, the method of designing multilevel visual features of images and applying them to image sentiment analysis has been widely tracked. Yanulevskaya et al. [15] first proposed low-level visual features, including Gabor and Wiccest features, to classify the emotions of artworks. Soli and Lenz [26] introduced an image descriptor based on color and emotion. This method is derived from psychophysical experiments for image classification and uses SIFT features for emotion prediction. Machajdik and Hanbury [2], based on art and psychological theories, defined a rich handcrafted middle-level feature from the aspects of composition, color change, and texture. Zhao et al. [16] introduced the middle-level visual features designed based on the concept of principle-of-art to extract emotion features (PAEF) to classify image emotion. However, compared with the features extracted from the CNN model, these manual features are mainly concentrated on low-level visual features. Due to the limited feature types and lack of exploration of high-level semantic information in images, it is difficult to cover all important factors related to image emotions.



(a)

(b)

Figure 1: Examples of emotion images and emotion arouse regions. (a) Images from the image emotion datasets. (b) Visualization of the main regions to invoke emotions.

In recent years, due to the excellent performance of CNN methods, researchers have applied the CNN method in image emotion analysis. Peng et al. [27] first applied the pretrained CNN model on ImageNet [28] for image sentiment analysis and achieved excellent classification results. You et al. [29] introduced a progressive strategy training to train the CNN model on a large-scale web image dataset to detect the emotion of the image. Rao et al. [17] proposed a multi-instance learning framework in order to obtain the multilevel deep representations of an image and obtained an exciting recognition result. You et al. [30] used the attention model to extract local emotional region features for emotional analysis. Yang et al. [31] proposed coupled CNN with two branches, which used both global and local information of an image. However, most of the studies did not fully use the local emotional regions of image, which limited the classification performance of the model.

2.2. Saliency Detection. Due to the powerful representation ability of deep features, the saliency detection method based on deep learning gradually surpasses the traditional method based on manual features [32–34]. Inspired by fully convolutional networks [35], more and more researches paid attention to predict the saliency map at the pixel level. Liu et al. [36] introduced an attention mechanism to guide the feature integration process by a U-shape model. Liu et al. [37] proposed a two-stage network algorithm. The algorithm generates a rough saliency map and combines local context information to refine the saliency map recursively and hierarchically. Hou et al. [38] introduced short connections in the multiscale side output to capture fine details. Zhang et al. [39] used a bidirectional structure to pass messages between the multilevel features extracted by the convolutional neural network to better predict the saliency map. Xiao et al. [40] first used a distracted detection network D-Net to crop the interference region in the image and then used the saliency detection network S-Net for saliency detection.

3. The Proposed Method

In order to improve image emotion recognition performance, an improved CNN is proposed, and the framework of our method is shown in Figure 2. The model includes the following improved components. (1) Two input branches: one is the original image input branch, and the other is the saliency image input branch. In the first branch, the network structure is modified based on Inception-v4 [41]. Firstly, the fully connected layer after the last convolutional layer in the Inception-v4 network is removed. Secondly, the side branch structure is introduced at three different depths of the network, and each side branch structure is composed of a convolutional layer and the convolution kernel size is 1×1. In the second branch, the network structure is also modified based on Inception-v4, and the fully connected layer after the last convolutional layer is removed. (2) Three fully connected layers work after the two branch inputs are completed. (3) A softmax layer generates the probability of each category and works after the fully connected layers.

In the input branch of the original image, the image semantic features on the global view are obtained from the last fully connected layer, and the feature maps from the multiple layers of the network are obtained from the side branches, and these feature maps are used as the input to calculate the Gram matrix. In the input branch of saliency map, the feature of local emotion region is extracted from the last convolution layer. Semantic features, local emotional features, and low-level visual features of the image are integrated into the hybrid representation features of image emotion classification. Finally, the hybrid representation

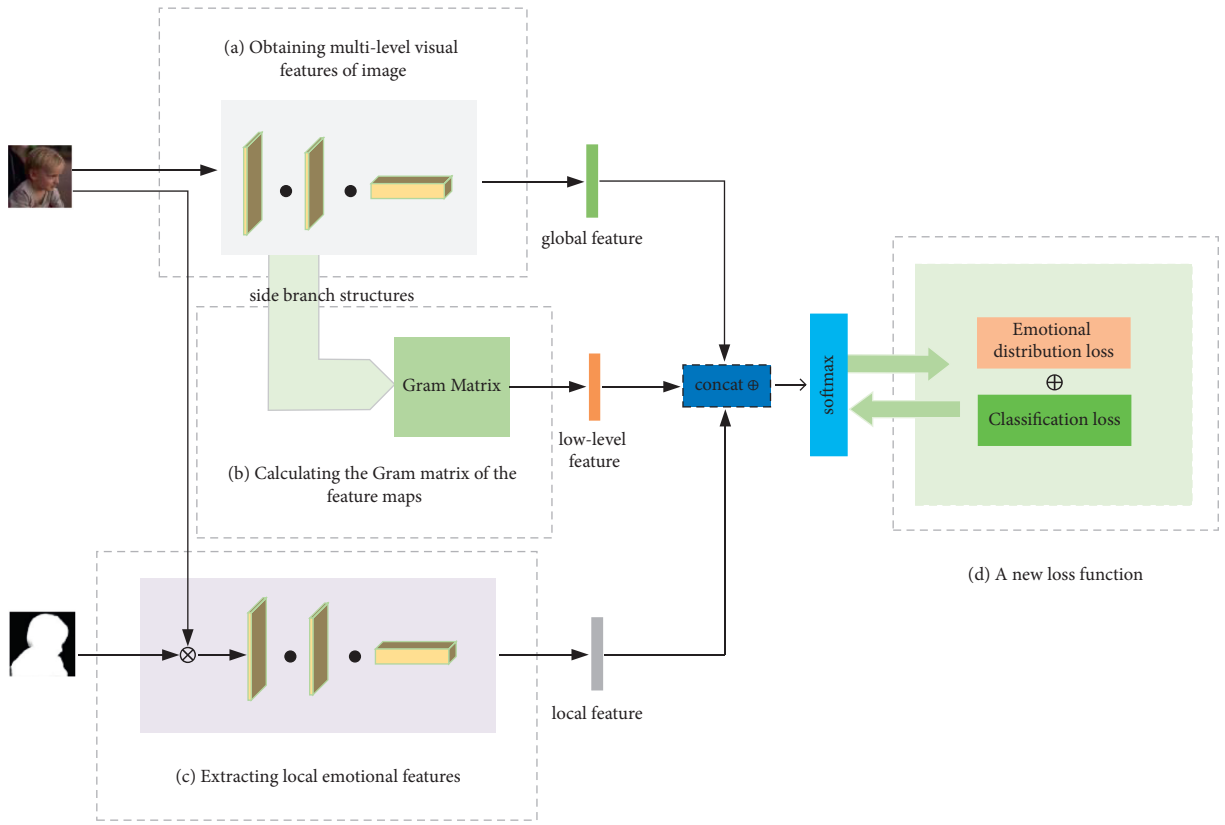


FIGURE 2: An overview of the proposed model. (a) The multilevel visual features are extracted by multiple side branch structures. (b) The Gram matrix of feature maps is calculated to reduce redundancy. (c) The saliency detection algorithm is used to locate the local emotion region of the image. (d) The hard label and probability label of image emotion category is used to design a new loss function.

features are input into the final fully connection layer and Softmax layer to predict the emotion category.

3.1. Saliency Detection and Local Emotional Features' Extraction. The human visual system only processes the vital part of image and meanwhile pays little attention to other parts, which prove that the human visual system has a certain mechanism to choose possible object positions when observing objects. So, the researchers consider that the object regions in the image are an emotional region with more emotions. In fact, the local regions covered by objects are more likely to attract people's attention and arouse their emotion. The saliency of the image highlights the degree of human attention to information-rich region and represents the different visual perceptions presented by different regions in the image. Based on the image saliency features, the saliency detection is used to locate the local region covered by objects in the image and extract the local emotional features of the image.

Firstly, image saliency detection algorithm is used to generate saliency image $Y = \{y_1, y_2, \ldots, y_n\}$, $y_i \in R^{w \times h}$, from corresponding original images $X = \{x_1, x_2, \ldots, x_n\}$, $x_i \in R^{3 \times w \times h}$, where w and h represent the width and height of the image, respectively. The saliency image is a binary image, and the size of the saliency image is the same as that of the original image. The element value of the object region of the original image is 1, while the element value of the nonobject region is 0. Thus, the local emotion region T can be calculated according to

$$T = X \bullet Y, \tag{1}$$

where $\bullet$ is the operator to multiply the elements of matrix X and the matrix Y. Then, input T into the saliency image input branch of the Siamese network to extract the local emotional features of the image.

3.2. Gram Matrix and Low-Level Visual Feature Extraction. The low-level visual features of the image are mainly concentrated in the shallow layers of the neural network [17]. There exists a problem of redundancy if we directly use the feature map obtained from the shallow layer of the network to describe the low-level visual features because the feature map also contains the image content information (e.g., objects and general scenery) [18].

In this paper, the low-level visual features are transformed by Gram matrix operation to reduce the redundancy. For each layer, use the feature maps to calculate the Gram matrix with the following steps. Firstly, vectorize each feature map F_i of size $w \times w$ in the convolutional layer to obtain a one-dimensional vector of length $L = w \times w$. Secondly, combine one-dimensional vectors in the order of the feature maps to obtain a matrix $F \in R^{N \times L}$, where N represents the number of feature maps in the convolutional layer.

Finally, calculate the Gram matrix $M \in R^{N\times N}$ of this convolutional layer according to

$$M = FF^T. \tag{2}$$

Each element M_{ij} in the Gram matrix is the inner product between the F_i and F_j, which can be obtained by

$$M_{ij} = \sum_k F_{ik}F_{jk}. \tag{3}$$

The procedure is summarized in Algorithm 1.

3.3. Loss Function of Emotional Subjectivity Constraint. In the collection of affective image data, the majority voting strategy is widely used to obtain the emotional label of the image. We calculate the distribution of image emotion based on the label probability to reduce the subjective influence of image emotion. The emotion theory research shows that the relationship between two emotions determines their similarity, and the two emotions from similar to completely opposite can be represented by Mikels' wheel [42]. As shown in Figure 3, a distance equation $\text{dist}(e_i, e_{i-1} = ''\text{feart}'')$ is defined in Mikels' wheel to quantify two emotional relationships. For example, the distance between the emotion fear and the emotion sadness is $\text{dist}(\text{fear}, \text{sadness}) = 1$, and the distance between the emotion fear and the emotion disgust is $\text{dist}(\text{fear}, \text{disgust}) = 2$, which indicates that the similarity between the emotion sadness and the emotion fear is higher.

Based on the definition of distance in Mikels' Wheel, the probability distribution of dominant emotion and other emotion can be calculated according to

$$f(i) = \begin{cases} \dfrac{\left(1/\text{dist}_{ij}\right)}{\sum_{i\neq j}\left(1/\text{dist}_{ij}\right)}\left(1 - p_j^*\right), & i \in V, \\ 0, & i \notin V, \end{cases} \tag{4}$$

where j is the dominant emotion category of the image, V denotes all the sentiment of the same polarity with the dominant emotion j, p_j^* is the probability of dominant emotion, and $f(i)$ is the probability of other emotions except the dominant emotion j. So, the probability distribution label of image emotion $d(i) = \{d_1, d_2, \ldots, d_n | n = 8\}$ can be obtained, and the sum of probabilities distribution $\sum d_i$ is normalized to 1.

Through using the hard label and probability distribution label, a new loss function can be designed according to

$$L_{\text{sub}j} = (1 - \lambda)L_{\text{cls}} + \lambda L_{kl}, \tag{5}$$

where L_{cls} is the cross-entropy classification loss, and it can be calculated by

$$L_{\text{cls}} = -\sum_i y_i \log(p_i), \tag{6}$$

where y_i is the ground truth label and p_i represents the probability that the image belongs to the i emotion category. Then, the Kullback–Leibler divergence [43] is used to measure the loss between probability distribution label $d(i)$ and predict emotion distribution p_i. Here, λ controls the weight of L_{kl}, and L_{kl} can be calculated by

$$L_{kl} = \sum_i d(i)\log(p_i). \tag{7}$$

4. Experiments and Results

In this section, our method is compared with other methods on FI, IAPSsubset, ArtPhoto, and Abstract datasets to evaluate our model.

4.1. Datasets. In the work of image emotion analysis, the widely used datasets mainly include FI, IAPSsubset, ArtPhoto, and Abstract, and the number of image samples in these datasets is shown in Table 1.

Flickr and Instagram (FI) [1]: this emotional dataset consists of about 23308 affective images. These pictures are collected by using 8 emotions as search keywords on Flickr and Instagram social networking sites. Then, these images were further labeled by Amazon Mechanical Turk, and the label of each image was done by five people voting.

In fact, the actual number of images that can be acquired in this dataset is 22,598 because the network connection for some images has failed. Table 2 shows the statistics of the number of available images.

IAPSsubset [2]: international affective image system (IAPS) is an international general emotion image dataset, which is widely used in image emotion classification. The dataset contains 1182 documentary-style natural images. Mikels et al. [42] selected 395 images from IAPs dataset and mapped them to eight emotion categories.

ArtPhoto [2]: in this dataset, photos are selected from the art photo-sharing website with emotion category as the search keyword, with a total of 806 photos. The emotional category of a photo is determined by the artist who uploaded it.

Abstract [2]: this dataset contains 228 abstract paintings. The emotional category of each abstract painting is decided by 14 different people. The emotion that gets the most votes is the emotion category of each image.

4.2. Implementation Details. The experiment was conducted on a computer based on the Pytorch environment. The computer used Intel(R) Xeon(R) CPU E5-2640 2.40 GHz CPU and NVIDA GeForce GTX TITAN GPU (12G memory). Our classification model is a Siamese network, and the backbone networks of the two branches are Inception-v4. The images in the dataset are randomly divided into training set (80%) and test set (20%): the training set totally has 18,078 images, and the test set totally has 4519 images. The image first scales the image in the range of [320, 480] based on the shortest side, then flips the image horizontally to obtain a mirror image, and then randomly crops 299×299 image blocks from the original image and the mirror image as the input of the model. We use the parameters pretrained on ImageNet to initialize the backbone

Input: feature map F_i of size $w \times w$
Output: Gram matrix $M \in R^{N \times N}$
Step 1: for each feature map F_i,

$$F_i = \begin{pmatrix} f_{11}^i, & f_{12}^i, & f_{1w}^i, \\ \vdots & \vdots & \vdots \\ f_{w1}^i, & f_{w2}^i, & f_{ww}^i \end{pmatrix}$$

vectorize F_i in convolution layer into a one-dimensional vector
$L_i = \left(f_{11}^i, f_{12}^i, \cdots f_{1w}^i, \cdots f_{w1}^i, f_{w2}^i, \cdots f_{ww}^i, \right)$, denoted as $L_1, L_2, \ldots, L_i$ $i = 1, 2, 3, \ldots N$;
Step 2: combine N one-dimensional vectors L_i into a matrix $\boldsymbol{F}$ in the order of the feature maps, denoted as $F \in R^{N \times L}$, $L = w \times w$.

$$F = \begin{pmatrix} f_{11}^1, & \cdots & f_{12}^1, & f_{1w}^1, & \cdots & f_{w1}^1, & f_{w2}^1, & \cdots & f_{ww}^1, \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{11}^N, & f_{12}^N, & \cdots & f_{1w}^N, & \cdots & f_{w1}^N, & f_{w2}^N, & \cdots & f_{ww}^N, \end{pmatrix}$$

Step 3: get the transposed matrix $F^T \in R^{N \times L}$ of matrix $F \in R^{N \times L}$, and compute the Gram matrix $M \in R^{N \times N}$ according to equation (3).

Algorithm 1: Procedure for applying the Gram matrix to convert the feature map.

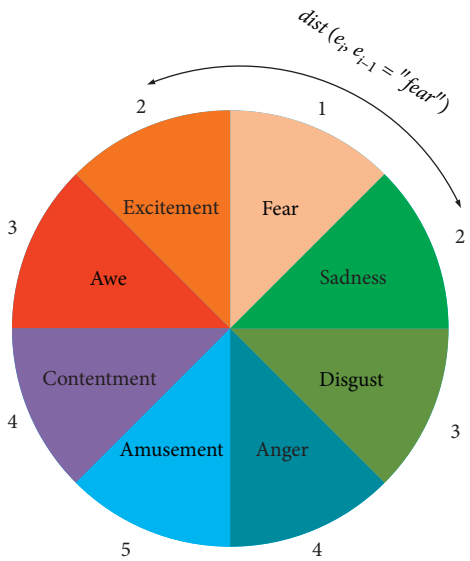


Figure 3: Mikels' emotion wheel and example of the emotion distance for emotion fear and other emotion [42].

Table 1: Statistics of the number of images in each image emotion datasets

Dataset	IAPSsubset	Artphoto	Abstract	FI
Amusement	37	101	25	4942
Anger	8	77	3	1266
Awe	54	102	15	3151
Contentment	63	70	63	5374
Disgust	74	70	18	1658
Excitement	55	105	36	2963
Fear	42	115	36	1032
Sadness	62	166	32	2922
Sum	**395**	**806**	**228**	**23308**

network of the model and use the stochastic gradient descent method to optimize the model. The parameters of our model are set as follows: the learning rate of model is set to 0.001,

Table 2: Statistics of the number of available images in FI emotion dataset.

Categories	Number of samples	Categories	Number of samples	Sum
Awe	3036	Disgust	1631	
Contentment	5268	Fear	1009	
Excitement	2808	Sadness	2771	
Amusement	4847	Anger	1228	
Sum	**15959**		**6639**	**22598**

and the weight decay is set to 0.0001. In particular, the learning rate is divided by 10 after every 5 epochs. The model is trained for up 20 epochs. The specific parameter settings are shown Table 3. Since the backbone network is a pre-trained model, the learning rate of the backbone network is set to 1/10 of the global learning rate for fine tuning.

4.3. Baseline

4.3.1. Handcrafted Features. In terms of handcrafted design features, GCH/LCH/GCH + BoW [44] used SIFT features based on bag-of-words to establish a 64-bit color histogram model for global color histogram (GCH) and local color histogram (LCH). Zhao et al. [16] introduced the middle-level visual features designed based on the concept of principles-of-art to extracted emotion features (PAEF) to classify image emotion. Rao et al. [45] proposed an emotion classification method based on multiscale blocks. Pyramid segmentation and simple linear iterative clustering (SLIC) method are used to segment the image into multiscale blocks. SentiBank [14] developed a visual sentiment ontology, which consist of 1200 concepts and associated classifiers, and each concept is composed of an adjective expressing emotion and a noun related to the object or scene.

4.3.2. Deep Features. In terms of deep features, AlexNet [8], VGG-16 [9], and Inception-v4 [41] all fine tune the pre-trained weights on the ImageNet dataset and complete the

Table 3: Initial parameters of our model.

Parameters	Value
Learning rate	0.001
Weight decay	0.0001
Momentum	0.9
Batch size	32
Epoch	20

emotion classification with the help of transfer learning. Deep SentiBank [46] proposed 2089-dim adjective-noun pair features based on CNN. PCNN [29] proposed a progressive strategy training to train the CNN model on the large-scale web image dataset to detect the emotion of the image. On the basis of AlexNet, Rao [17] obtained multilevel deep features by constructing multiple side branches in the network. Yang [47] proposed a learning method based on label distribution, which aims to solve the subjective problem of image emotion. WSCNet [31] proposed a weakly supervised coupled convolutional network with two branches.

4.4. Experimental Validation. In this paper, the classification model for large-scale emotional image dataset (FI) is initialized by using the parameters pretrained on the ImageNet dataset and then fine tuning the model on the FI dataset to complete the classification task. For small-scale datasets (IAPSsubset, Artphoto, and Abstract), the classification model is initialized by using the parameters pretrained on the FI dataset and then further fine tuning the model to complete the classification tasks.

4.4.1. The Effectiveness of Local Emotional Feature. To validate the effectiveness of the local emotional features, we designed a comparative experiment on the FI dataset. (1) Our model only uses the global feature from the last convolutional layer of the original image input branch of our model and low-level visual features. (2) Our model only uses the local emotional feature extracted from the local emotional region of the image. (3) Our model uses hybrid classification features composed of global semantic features, local emotional features, and low-level visual features. Table 4 shows the classification performance of our model with the three configurations on the FI dataset. Specifically, the global view only means that the model uses the global semantic feature and the low-level visual features, the emotional region only means that the model only uses the local emotional feature extracted from the local emotional region of the image, and the global view + emotional region means that the model uses hybrid classification features composed of global semantic features, local emotional features, and low-level visual features. As shown in Table 4, the model in (1) only uses global semantic features and low-level visual features, while the model in (3) uses local emotional features as supplementary information, and the classification accuracy of the model is improved about 4%, which shows that combining emotional features from local emotional regions can effectively improve emotional classification performance

Table 4: The classification accuracy on the FI dataset.

Method	Accuracy (%)
Global view only	66.14
Emotional region only	59.82
Global view + emotional region (ours)	**70.23**

than using global features only. In (2), when the model only uses the features from the local emotional region, the classification performance of the model is severely reduced, which illustrates the importance of extracting semantic features from the global view of the image.

In Figure 4, the classification confusion matrixes of our model are shown in the two configurations of whether or not to use image local emotional features. It can be seen that applying local emotional features can enhance the classification performance of model and produce a more balanced recognition result for each emotion category.

4.4.2. The Effectiveness of Gram Matrix Transform. In order to get more low-level visual features, we introduce multiple side branches into the network. Each side branch is composed of a convolution layer. We apply Algorithm 1 to each side branch, respectively, and transform the feature map to obtain the low-level visual feature of the image $\{G_i | i = 1, 2, 3, \ldots\}$. As shown in Table 5, C represents a hybrid feature composed of global semantic features and local emotional features, L represents the low-level visual features described by the feature map directly, and G represents the low-level visual features captured from the feature map by using the Gram matrix. In Table 5, the best classification result can be obtained by combining the feature C and feature $\{G_1, G_2, G_3\}$. The low-level visual features captured from the feature map can get better classification results. It also can be seen that when L_4, L_5 or G_4, G_5 from the high layers of the network are added, the classification accuracies decreased. Adding feature G_4, G_5 has less effect on classification performance compared with adding feature L_4 and L_5. This shows that the Gram matrix transform can effectively reduce the redundancy of image content information in the feature map.

4.4.3. The Effectiveness of Loss Function. Our new loss function $L_{\text{sub}j}$ is designed by using the hard labels and probability labels of image emotion categories, trying to reduce the impactive of image emotion subjectivity. Different from the cross-entropy loss function L_{cls}, $L_{\text{sub}j}$ maximizes the difference between emotion classes and emphasizes the relationship between emotion categories by comprehensively constraining the classification loss and emotion distribution loss. The two loss functions mentioned above were used to conduct comparative experimental on the FI dataset, and the results are shown in Table 6. As can be seen, the classification performance of the model has been improved after applying the $L_{\text{sub}j}$ loss function. In particular, the classification accuracy of our model is improved by about 1.4% after applying the $L_{\text{sub}j}$ loss function, which shows the effectiveness of our loss function.

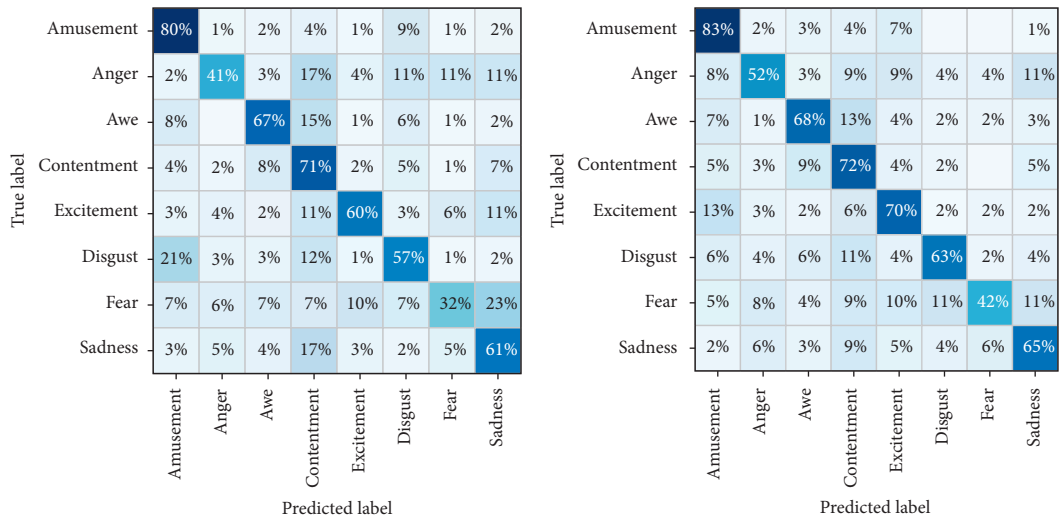


Figure 4: Classification confusion matrix on the FI dataset. (a) Our model without local emotional feature. (b) Our model with local emotional feature.

Table 5: Comparison of classification results on the FI dataset using different feature combinations.

Method	Accuracy (%)
C	67.8
$C + \{L_1\}$	68.20
$C + \{L_1, L_2\}$	68.67
$C + \{L_1, L_2, L_3\}$	69.12
$C + \{L_1, L_2, L_3, L_4\}$	67.52
$C + \{L_1, L_2, L_3, L_4, L_5\}$	67.13
C	67.8
$C + \{G_1\}$	68.54
$C + \{G_1, G_2\}$	69.30
$C + \{G_1, G_2, G_3\}$	70.23
$C + \{G_1, G_2, G_3, G_4\}$	68.74
$C + \{G_1, G_2, G_3, G_4, G_5\}$	68.31

Table 6: Comparison of classification results on the FI dataset using different loss functions.

Method	Accuracy (%)
AlexNet + L_{cls}	58.61
ResNet101 + L_{cls}	60.82
Inception-v4 + L_{cls}	60.75
Ours + L_{cls}	70.23
AlexNet + L_{subj}	60.32
ResNet101 + L_{subj}	62.77
Inception-v4 + L_{subj}	62.66
Ours + L_{subj}	**71.65**

4.4.4. The Choice of Parameter λ. In this work, parameter λ is used to control the weight of classification loss and sentiment distribution loss. When λ is set to 0, the proposed loss function is the cross-entropy loss, and λ is set to 1 and indicates that the proposed loss function is equal to KL loss essentially. Figure 5 shows the accuracy change under different values of parameterλ. When λ increases from 0 to 0.4, the classification performance has a significant improvement. However, when it further increases to more than 0.5, the classification accuracy begins to decrease. Figure 5 shows that when the weight of L_{ed} is set too large, it may lead too much ambiguity.

4.5. Compare with the Other Methods

4.5.1. Compare on Large-Scale Datasets. To further indicate the effectiveness of the proposed model, we compare it with the methods shown in Table 7. Our model has obviously achieved better results compared with the method based on manual features of SentiBank [14] through using hybrid representation features, which consist of global sematic, local visual, and low-level visual features. We can see that the performance of our model is better than those of CNN networks specifically proposed for object recognition tasks in Table 7, such as AlexNet [8], VGG-19 [9], and Inception-v4 [41]. Moreover, our model achieves better classification performance compared with the deep learning model proposed for image emotion classification, such as Yang et al. [47], MldrNet [17], and WSCNet [31], which shows the

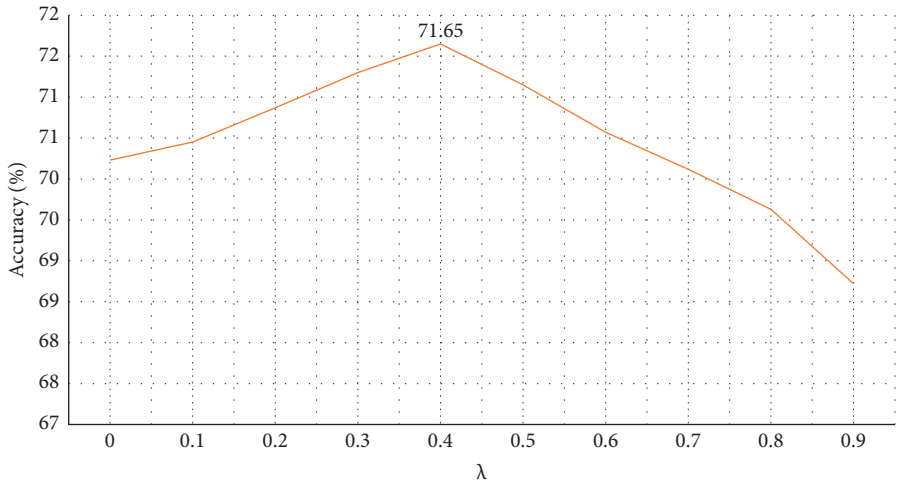


Figure 5: Impact of different λ on the FI dataset.

Table 7: Comparison of classification performance on the FI dataset.

Method	Accuracy (%)
AlexNet [8]	58.61
VGG-16 [9]	59.75
Inception-v4 [41]	60.75
MldrNet [17]	67.75
Yang [47]	67.64
WSCNet [31]	70.07
Ours	**71.65**

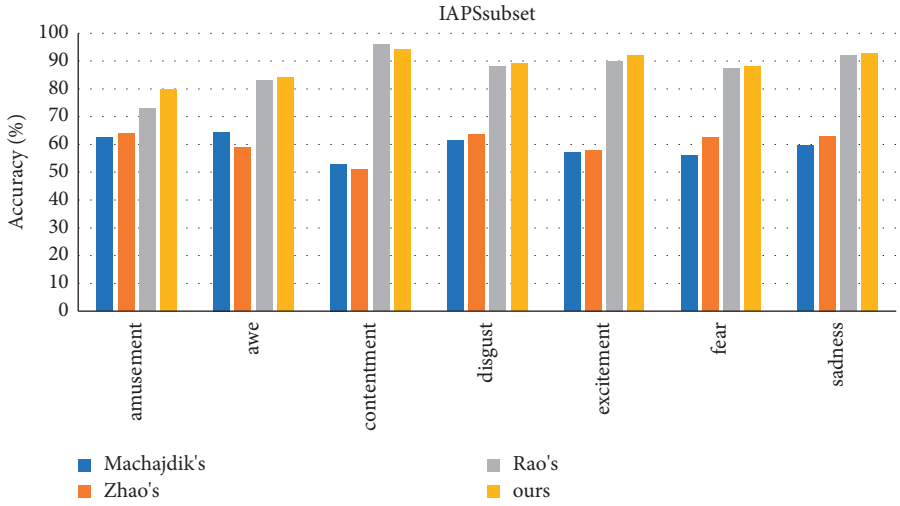


Figure 6: Performance evaluation on the IAPSsubset dataset.

effectiveness of our global and local hybrid representation features, as well as the effectiveness of our loss function.

4.5.2. Compare on Small-Scale Datasets. In order to verify the performance of the model more comprehensively, we also designed a comparative experiment on a small dataset, including IAPSsubset, Abstract, and ArtPhoto. Before the experiment, we randomly divided the image samples of each category in the dataset into 5 batches. Then, 5-fold cross validation is performed to obtain results. Especially, the emotion category anger has only 8 and 3 samples in the Abstract and IAPSsubset datasets, respectively, performing 5-fold cross validation is not enough. Therefore, the classification result of emotion anger on these two datasets is not reported. The experiment results are shown in Figures 6–8.

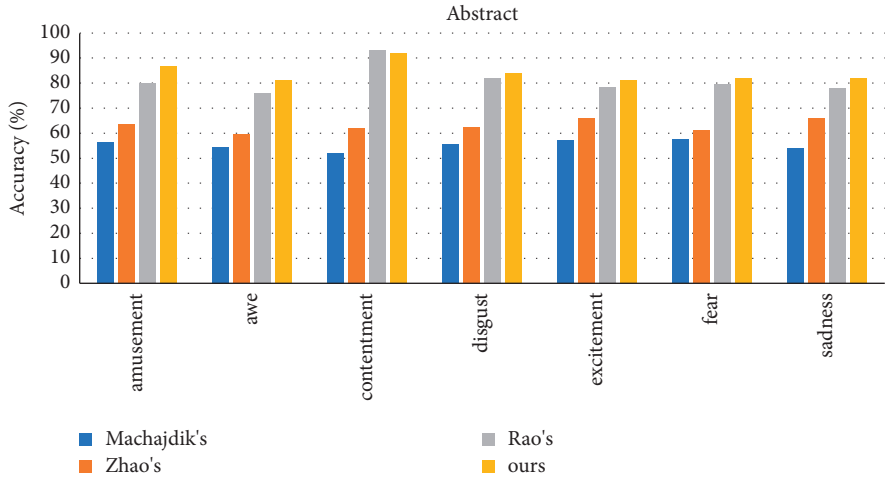


Figure 7: Performance evaluation on the Abstract dataset.

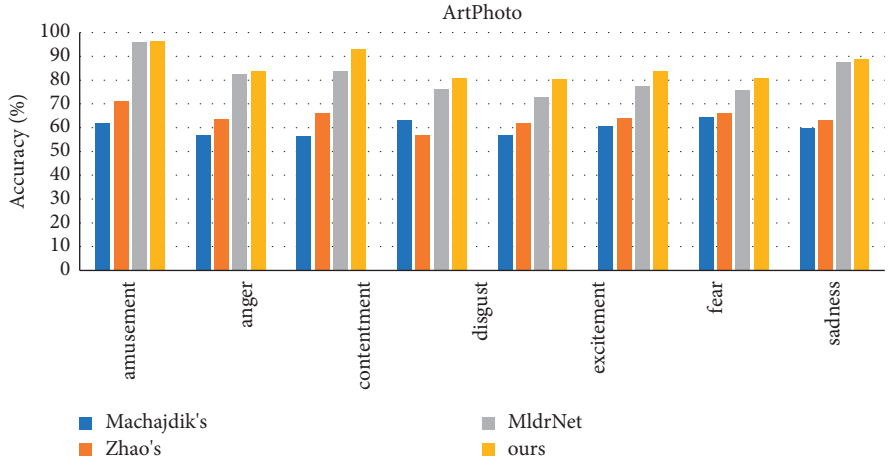


Figure 8: Performance evaluation on the ArtPhoto dataset.

Our method outperforms to Machajdik et al. [2], Zhao et al. [16], and MldrNet [17] in IAPSsubset, Abstract, and Artphoto.

5. Conclusions

In this paper, a CNN framework based on saliency detection and Gram matrix is proposed to improve image emotion recognition performance, and our method have been applied on many famous problems, including FI (Flickr and Instagram), IAPSsubset, ArtPhoto, and Abstract. The classification accuracies have been compared with those of other competing methods in the literatures, and the results show that our method has improved the image emotion recognition performance. Through experimental analyzing, it can be drawn that saliency detection, Gram matrix transformation, and new loss function are effective in increasing recognition accuracy, which indicates that the proposed method has potential application ability. In the future work, our main task is to integrate this improved CNN into the actual applications and conduct emotion recognition for video data automatically to better serve the society.

Data Availability

The datasets used in this study are Flickr and Instagram (FI) (https://onedrive.live.com/?authkey=%21AH57YMUbsP%2DqNls&cid=AB6522E29F6ED9A0&id=AB6522E29F6ED9A0%21101730&parId=AB6522E29F6ED9A0%21101729&action=defaultclick), Abstract (https://www.imageemotion.org/testImages_abstract.zip), IAPSsubset (https://www.csea.phhp.ufl.edu/media.html), and ArtPhoto(https://www.imageemotion.org/testImages_artphoto.zip).

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China, under Grant no.61977018, the Research Foundation of Education Bureau of Hunan Province of China, under Grant no. 16B006, the Hunan Provincial Natural Science Foundation of China, under Grant no.

2020JJ4626, and the Scientific Research Fund of Hunan Provincial Education Department of China, under Grant no. 19B004.

References

[1] 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*, pp. 308–314, Palo Alto, CA, USA, May 2016.

[2] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 83–92, Firenze, Italy, October 2010.

[3] Y.-H. Chew, L.-K. Wong, J. See, H.-Q. Khor, and B. Abivishaq, "LiteEmo: lightweight deep neural networks for image emotion recognition," in *Proceedings of the 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, Kuala Lumpur, Malaysia, September 2019.

[4] W. Wang, Y. Yu, and J. Zhang, "Image emotional classification: static vs. dynamic," in *Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, pp. 6407–6411, Hague, Netherlands, October 2004.

[5] D. T. Priya and J. D. Udayan, "Affective emotion classification using feature vector of image based on visual concepts," *The International Journal of Electrical Engineering & Education*, vol. 60, 2020.

[6] X. He and W. Zhang, "Emotion recognition by assisted learning with convolutional neural networks," *Neurocomputing*, vol. 291, pp. 187–194, 2018.

[7] B. Chen, W. Tan, G. Coatrieux, Y. Zheng, and Y. Q. Shi, "A serial image copy-move forgery localization scheme with source/target distinguishment," *IEEE Transactions on Multimedia*, vol. 20, p. 1, 2020.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[10] Y. Qiao, Y. Tian, Y. Liu, and J. Jiao, "Genetic feature fusion for object skeleton detection," *Security and Communication Networks*, vol. 2021, Article ID 6621760, 9 pages, 2021.

[11] Y. Wang, X. Cui, Z. Gao, and B. Gan, "Fed-scnn: a federated shallow-cnn recognition framework for distracted driving," *Security and Communication Networks*, vol. 2020, Article ID 6626471, 10 pages, 2020.

[12] B. Chen, X. Ju, B. Xiao, W. Ding, Y. Zheng, and V. H. C. De Albuquerque, "Locally GAN-generated face detection based on an improved Xception," *Information Sciences*, vol. 572, pp. 16–28, 2021.

[13] F. Cen, X. Zhao, W. Li, and G. Wang, "Deep feature augmentation for occluded image classification," *Pattern Recognition*, vol. 111, Article ID 107737, 2021.

[14] D. Borth, T. Chen, R. Ji, and S. F. Chang, "Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content," in *Proceedings of the 21st ACM International Conference on Multimedia*, pp. 459-460, Barcelona, Spain, October 2013.

[15] V. Yanulevskaya, J. C. Gemert, K. Roth, A. K. Herbold, N. Sebe et al., "Emotional valence categorization using holistic image features," in *Proceedings of the 2008 15th IEEE International Conference on Image Processing*, pp. 101–104, San Diego, CA, USA, October 2008.

[16] 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*, pp. 47–56, Orlando, FL, USA, November 2014.

[17] T. Rao, 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.

[18] X. Alameda-Pineda, E. Ricci, Y. Yan, and N. Sebe, "Recognizing emotions from abstract paintings using non-linear matrix completion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5240–5248, San Francisco, CA, USA, August 2016.

[19] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2414–2423, San Francisco, CA, USA, August 2016.

[20] M. Elad and P. Milanfar, "Style transfer via texture synthesis," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2338–2351, 2017.

[21] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer," 2017, https://arxiv.org/abs/1701.01036.

[22] T. Rao, X. Li, H. Zhang, and M. Xu, "Multi-level region-based convolutional neural network for image emotion classification," *Neurocomputing*, vol. 333, pp. 429–439, 2019.

[23] F. Zhou, C. Cao, T. Zhong, and J. Geng, "Learning meta-knowledge for few-shot image emotion recognition," *Expert Systems with Applications*, vol. 168, Article ID 114274, 2021.

[24] D. T. Priya and J. D. Udayan, "Transfer learning techniques for emotion classification on visual features of images in the deep learning network," *International Journal of Speech Technology*, vol. 23, no. 2, pp. 361–372, 2020.

[25] X. Liu, N. Li, and Y. Xia, "Affective image classification by jointly using interpretable art features and semantic annotations," *Journal of Visual Communication and Image Representation*, vol. 58, pp. 576–588, 2019.

[26] M. Solli and R. Lenz, "Color based bags-of-emotions," in *Proceedings of the International Conference on Computer Analysis of Images and Patterns*, pp. 573–580, Münster, Germany, September 2009.

[27] KC. Peng, T. Chen, A. Sadovnik, and A. Gallagher, "A mixed bag of emotions: model, predict, and transfer emotion distributions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 860–868, Boston, MA, USA, June 2015.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "Imagenet: a large-scale hierarchical image database," in *Proceedings of the I2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Miami, FL, USA, June 2009.

[29] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, Austin, Texas, January 2015.

[30] Q. You, H. Jin, and J. Luo, "Visual sentiment analysis by attending on local image regions," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 231–237, San Francisco, CA, USA, February 2017.

[31] J. Yang, D. She, Y.-K. Lai, P. L. Rosin, and M.-H. Yang, "Weakly supervised coupled networks for visual sentiment

analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7584–7592, San Francisco, CA, USA, August 2018.

[32] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: a discriminative regional feature integration approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2083–2090, San Francisco, CA, USA, August 2013.

[33] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep CNN features," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5012–5024, 2016.

[34] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: contrast based filtering for salient region detection," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 733–740, Providence, RI, USA, June 2012.

[35] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, San Francisco, CA, USA, August 2015.

[36] N. Liu, J. Han, and M. H. Yang, "Picanet: learning pixel-wise contextual attention for saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3089–3098, San Francisco, CA, USA, August 2018.

[37] N. Liu and J. Han, "Dhsnet: deep hierarchical saliency network for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 678–686, San Francisco, CA, USA, August 2016.

[38] Q. Hou, M. -M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3203–3212, San Francisco, CA, USA, August 2017.

[39] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1741–1750, San Francisco, CA, USA, August 2018.

[40] H. Xiao, J. Feng, Y. Wei, M. Zhang, and S. Yan, "Deep salient object detection with dense connections and distraction diagnosis," *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3239–3251, 2018.

[41] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, CA, USA, February 2017.

[42] 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*, vol. 19, no. 3, pp. 632–645, 2016.

[43] F. Pérez-Cruz, "Kullback-Leibler divergence estimation of continuous distributions," in *Proceedings of the 2008 IEEE International Symposium on Information Theory*, pp. 1666–1670, Toronto, Canada, July 2008.

[44] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, "Analyzing and predicting sentiment of images on the social web," in *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 715–718, Firenze, Italy, October 2010.

[45] T. Rao, M. Xu, H. Liu, J. Wang, and I. Burnett, "Multi-scale blocks based image emotion classification using multiple instance learning," in *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP)*, pp. 634–638, Phoenix, AZ, USA, August 2016.

[46] T. Chen, D. Borth, T. Darrell, and S. F. Chang, "Deepsentibank: visual sentiment concept classification with deep convolutional neural networks," 2014, https://arxiv.org/abs/1410.8586.

[47] J. Yang, D. She, and M. Sun, "Joint image emotion classification and distribution learning via deep convolutional neural network," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 3266–3272, Melbourne, Australia, August 2017.