Color Enhanced Cross Correlation Net for Image Sentiment Analysis

Shulan Ruan, Kun Zhang, *Member, IEEE*, Le Wu *Member, IEEE*, Tong Xu, *Member, IEEE*, Qi Liu, *Member, IEEE*, and Enhong Chen, *Senior Member, IEEE*

***Abstract*—Automatic analysis of image sentiment has gained considerable attention with the increasing throughput of user-generated visual contents online. Recently, researchers generally tend to design different Convolutional Neural Networks (CNNs) to extract image content features for sentiment analysis. However, they underestimated the importance of image color, which has been proved very crucial for image sentiment expressing by psychology and art theory. Moreover, we further observe that the coordination of content and color is the main form of image sentiment expressing. Different combinations of content and color could express extremely different sentiments. To that end, in this paper, we propose a Color Enhanced Cross Correlation Net (CECCN), a novel architecture for image sentiment analysis that not only leverages contents and colors simultaneously, but also takes their correlations into consideration. Specifically, we first use a pre-trained CNN to extract content features and color moment to collect color features from multiple color spaces. Then, we propose a novel Cross Correlation (CC) method to model the correlations between content features and color features with attention mechanism and sequence convolution, in which sentiment expressing of content and color can be enhanced by each other. Finally, we integrate these two types of information for better image sentiment analysis. Extensive experiments on two popular and well-studied benchmark datasets demonstrate the superiority and rationality of our proposed CECCN.**

***Index Terms*—Image sentiment analysis, neural network, cross correlation, feature representation.**

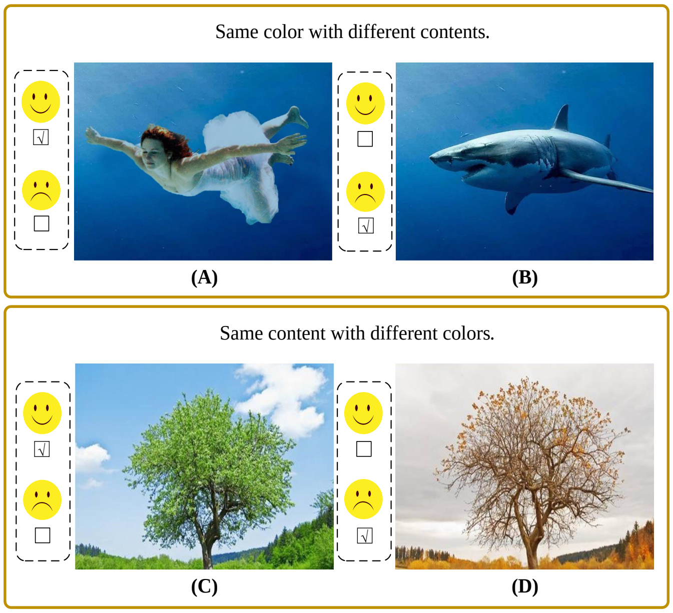

Same color with different contents.
√
□
□
√
(A)
(B)
Same content with different colors.
√
□
□
√
(C)
(D)

Fig. 1. (A) and (B) reflect the importance of image content. (C) and (D) are examples for images consisting of similar content but with different colors, which may indicate different sentiments.

I. Introduction

IMAGE Sentiment Analysis aims to automatically figure out sentiments from images. It has broad applications in many areas, such as Opinion Mining [1], Image Retrieval [2] and Recommender System [3]. Moreover, with the increasing popularity of social networks, more and more people tend to express their feelings and opinions on the Internet with visual contents. Thus, this task has become a hot topic and plenty of efforts in this area have been made to help to understand user behaviors [4], [5].

Much progress has been made in this area with the successful accomplishments of deep learning [6] and publication of large scale datasets [7]. Most recent studies tend to utilize pre-trained CNNs to automatically extract the features of images for image sentiment analysis. For example, Campos et al. [8] applied CNN to image sentiment prediction through fine-tuning experiments. Song et al. [9] integrated visual attention into CNN framework to find important areas affecting sentiment. Other researchers focused on the usage of local information of sentiment analysis. Wu et al. [10] demonstrated that reasonably combining local objects and global image could improve the performance for image sentiment analysis.

However, most of these works focused on designing various CNNs to extract image features for precise image sentiment analysis. They focused more on the image content and underestimated the importance of colors when expressing sentiments with images, which have been proven very crucial by psychology knowledge [11], [12], [13] and art theory [14]. For example, Jacobs et al. [15] observed that long-wavelength colors (e.g., red and yellow) are more arousing than short-wavelength colors (e.g., blue and green), and artists usually leverage colors to express certain emotions [14]. Moreover, the interaction between image color and image content is still unclear. Previous methods have weaknesses in revealing how image color affects the sentiment of images. Taking Fig. 1 (A) and (B) as examples, these two images have similar colors but extremely different contents. Obviously, Fig. 1 (A) looks content and excited based on the content of a beautiful girl diving in the sea, while the shark in Fig. 1 (B) conveys

S. Ruan, T. Xu, Q. Liu and E. Chen are with the Anhui Province Key Laboratory of Big Data Analysis and Application, School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China. (email: slruan@mail.ustc.edu.cn, tongxu, qiliuql, cheneh@ustc.edu.cn).

K. Zhang and L. Wu are with School of Computer and Information, Hefei University of Technology, Hefei 230029, China. (email: zhang1028kun, lewu.ustc@gmail.com).

the sentiment of fear to us. We could observe that different contents will change the sentiment attributes of the same color. Correspondingly, in (C) and (D), image contents are almost the same while colors are completely different. Fig. 1 (C) gives us a positive and vital feeling. On the contrary, Fig. 1 (D) shows a negative and desolate state. Based on this example, we could obtain that different colors will change the sentiment attributes of the same content. Therefore, the interactions (coordination and correlation) between image color and content are extremely important for image sentiment analysis. How to analyze the coordination and correlation of image color and content to further improve model performance is the main focus of this paper.

Unfortunately, there are still many unique challenges inherent in designing an effective solution to integrate image color and content information in a deep learning framework for image sentiment analysis. First of all, it is difficult to represent rich color information concerning image sentiment in a deep learning framework. Traditional methods [16] usually directly extract a certain color space features as a latent vector with a simple statistical method (e.g., color histogram), while the single low-dimension vector may not be able to fully represent color information and reveal the complex relation between color and sentiment. Moreover, deep learning methods (e.g., CNNs) process the color and content of images in a unified manner and leverage the pixel value to represent the mixture of color and content. To this end, rich color information (e.g., interaction with content, global color distribution) is underestimated. Secondly, the coordination and correlation of image content and image color is still unclear. How to utilize their coordination and correlation to enhance image sentiment analysis is still a great challenge that remains unsolved.

To address the challenges mentioned above, in this paper, we propose a Color Enhanced Cross Correlation Net (CECCN), a novel architecture that takes content and color into consideration simultaneously for better image sentiment analysis. To be specific, we first utilize a pre-trained CNN to pay more attention to content features. While for color information, we leverage color moment method to collect color features (i.e., mean, variance and skewness) from multiple spaces (e.g., RGB, HSV). Then, we employ a color embedding method to enrich the sentiment information carried by colors. In order to better model the correlations between content and color on image sentiment and enhance their feature representations, we design a novel *Cross Correlation (CC)* method to model their correlations with each other. Along this line, content features can be better employed to choose the most informative color features with attention mechanism, and color features are fully utilized to enrich content feature representations via sequence convolution. Finally, we fuse these two well-learned features with a weighted sum method for better image sentiment analysis.

As an emphasis, main contributions of our work can be concluded as follows:

- We observe the importance of complex interaction between image color and content for image sentiment expressing, and propose to leverage the image color to enhance the image sentiment analysis.
- We propose a novel CECCN method which takes both color and content into consideration. Moreover, we design a Cross Correlation method to model their correlations, thus enhance their feature representations respecting image sentiment.
- Extensive experiments on two popular and well-studied benchmark datasets and two different classification tasks demonstrate the superiority and rationality of our proposed method compared with the baseline methods.

The remainder of this paper is organized as follows. In Section II, we introduce the related work. Then, the model and technical details we propose are presented in Section III. In Section IV, we conduct various experiments on public datasets and give detailed analysis for convincing interpretability. Finally, we conclude our work and describe future work in Section V.

II. Related work

In this section, we will review the related work on image sentiment analysis and effects of color on image sentiment that are closely related to this paper.

A. Image Sentiment Analysis

1) Shallow modeling methods: In the earlier attempt of image sentiment analysis, most shallow modeling methods on image sentiment analysis often employed low-level hand-craft features. Stefan et al. [16] proposed to predict the sentiment of images using pixel-level features. Inspired by psychology and art theory, Machajdik et al. [14] exploited theoretical and empirical concepts to extract image features that are specific to the domain of artworks with emotional expressions, such as color and texture. Similarly, Lu et al. [17] investigated how shape features in natural images influence emotions aroused in human beings. Zhao et al. [18] made use of the concept of principles-of-art and its influence on image sentiment. With the establishment of large-scale Visual Sentiment Ontology (VSO), SentiBank [19] was proposed to detect the presence of 1,200 ANPs in an image for visual sentiment analysis. These shallow modeling methods with carefully designed sentiment-related hand-craft features have been proved to be effective and attracted a lot of research attention for a long time before the popularity of various deep modeling methods.

2) Deep modeling methods: With the successful accomplishments of deep learning and publication of large scale datasets [7], CNN has achieved impressive performance in various computer vision tasks, such as object classification [6], image caption [20], [21] and visual question answering [22], [23]. With the help of pre-trained models, many recent studies tried to introduce CNN for image sentiment analysis [24], by training models on large scale datasets first and then fine-tuning on their own datasets to better extract image content features. You et al. [25] utilized a progressive strategy to train CNN on their dataset. Campos et al. [8] explored how CNN could be specifically applied to the task of visual sentiment prediction through fine-tuning experiments and rigorous architecture analysis. Song et al. [9] presented SentiNet-A, a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3118208, IEEE Transactions on Multimedia

novel architecture that integrates visual attention into CNN sentiment classification framework, by jointly learning with multi-scale saliency detection in different CNN layers. In [26], Zhang et al. demonstrated that deep models mainly rely on the image content but miss the image style information. To this end, they proposed a novel CNN model that learned and integrated the content information from the high layers of the deep network with the style information from the lower layers. Yang et al. [27] proposed a deep framework for automatically discovering the affective regions of images, and built an image sentiment prediction model using a deep CNN, which utilized the holistic and local information from both the global image and the local regions. Similarly, Wu et al. [10] demonstrated that reasonably utilizing the local information could improve the performance for image sentiment analysis. In [28], She et al. were dedicated to automatically selecting relevant soft proposals given weak annotations (e.g., global image labels), thereby significantly reducing the annotation burden. By applying the residual attention model, RA-DLNet [29] was proposed to focus on crucial sentiment-rich, local regions in the image. Considering the image sentiment is usually closely related to humans appearing in the image, Zheng et al. [30] proposed to model the contribution of human faces as a special local region for sentiment prediction. Based on Bayesian network, Zhang [31] proposed an object semantics sentiment correlation model to leverage the influence of object semantics on image sentiment analysis.

3) Multimodal methods: With the popularity of social platforms, many users often use images and text to record and share their daily lives and moods on the Internet. Therefore, associated with shared images, a large amount of textual metadata including image titles, image tags, and text descriptions have also become available. In recent years, some researchers have made efforts on how to utilize both raw image and textual metadata to boost visual sentiment analysis performance. Katsurai et al. [32] proposed a novel approach that exploited latent correlations among multiple views: visual and textual views, and a sentiment view constructed using SentiWordNet. For some images such as advertisement posters, words or sentences are added on the image with image-editing software, Felicetti et al. [33] argued that it is essential to not only analyze the sentiment of the visual elements but also to correctly understand the meaning of the included text and to analyze it accordingly. They performed multimodal sentiment analysis by extracting both visual features and text features with the OCR model [34]. In [35], Zhu et al. demonstrated that visual and textual information should differ in their contribution to sentiment analysis. Their proposed model learns a robust joint visual-textual representation by incorporating a cross-modality attention mechanism and semantic embedding learning based on bidirectional recurrent neural network. Besides the subjective text metadata, Ortis et al. [36] also exploited objective text description of images for visual sentiment analysis. Many other methods have also contributed to multimodal sentiment analysis and cross-domain information process [37], [38], [39].

In this paper, we focus on only using the raw image to analyze the image sentiment without textual metadata. Although these deep modeling methods that employ pre-trained CNNs to extract image content-related features have become the mainstream in recent years, most of these models overlooked a very important phenomenon that color, a kind of low-level image feature directly related to human vision, plays a very important role in image sentiment analysis. Therefore, in this paper, we integrate color information into a deep learning framework and pour attention to model the complex interaction between image color and content for image sentiment expressing.

B. Effects of Color on Image Sentiment

Little work pays close attention to the use of color factor for image sentiment analysis recently in literature. Actually, art theory and psychology studies have been investigated on the effects of color for image sentiment. In art theory, colors are usually effectively used by artists to induce emotional effects [14].

Extensive psychological studies also have been carried out to confirm the importance of color to image sentiment expressing. Experimental studies that have used physiological measures (e.g., galvanic skin response, electroencephalograph) generally have shown that red and yellow were indeed more arousing than blue and green [11], [12]. Profusek et al. [40] investigated the effects of rooms painted in red versus Baker-Miller pink on state anxiety. As hypothesized, pink elicited less anxiety than red. Jacobs et al. [15] investigated the effects of four primary colors (i.e., red, yellow, green, blue), projected onto a large screen. They further observed that long-wavelength colors (e.g., red and yellow) are more arousing than short-wavelength colors (e.g., blue and green). In [41], Valdez et al. conducted various regression analyses to test for possible relationship between brightness and image sentiment. They drew the conclusion that relationship was very strong and highly significant. Besides, Guilford et al. [13] also found that brighter and more saturated colors elicited greater pleasure, with the relationships tending to be curvilinear.

The results of these studies intuitively demonstrated that color affects sentiment expressing and judgment. We could make better image sentiment analysis, if color factor could be taken into account.

III. Model structure

In this section, we mainly introduce the technical details of our proposed Color Enhanced Cross Correlation Net (CECCN), a novel dual-branch deep method for image sentiment analysis.

The overall architecture is shown in Fig. 2, which consists of three components: 1) *Feature Representation*, extracting image content features and color features with CNN and color moment, respectively; 2) *Cross Correlation*, cross correlating the two different feature representations with attention mechanism and sequence convolution for information enhancement; 3) *Label Prediction*, utilizing the two representations to predict the sentiment classification results robustly.

A. Feature Representation

In this component, we extract content-related and color-related features simultaneously since both image content and image color are crucial for image sentiment analysis.

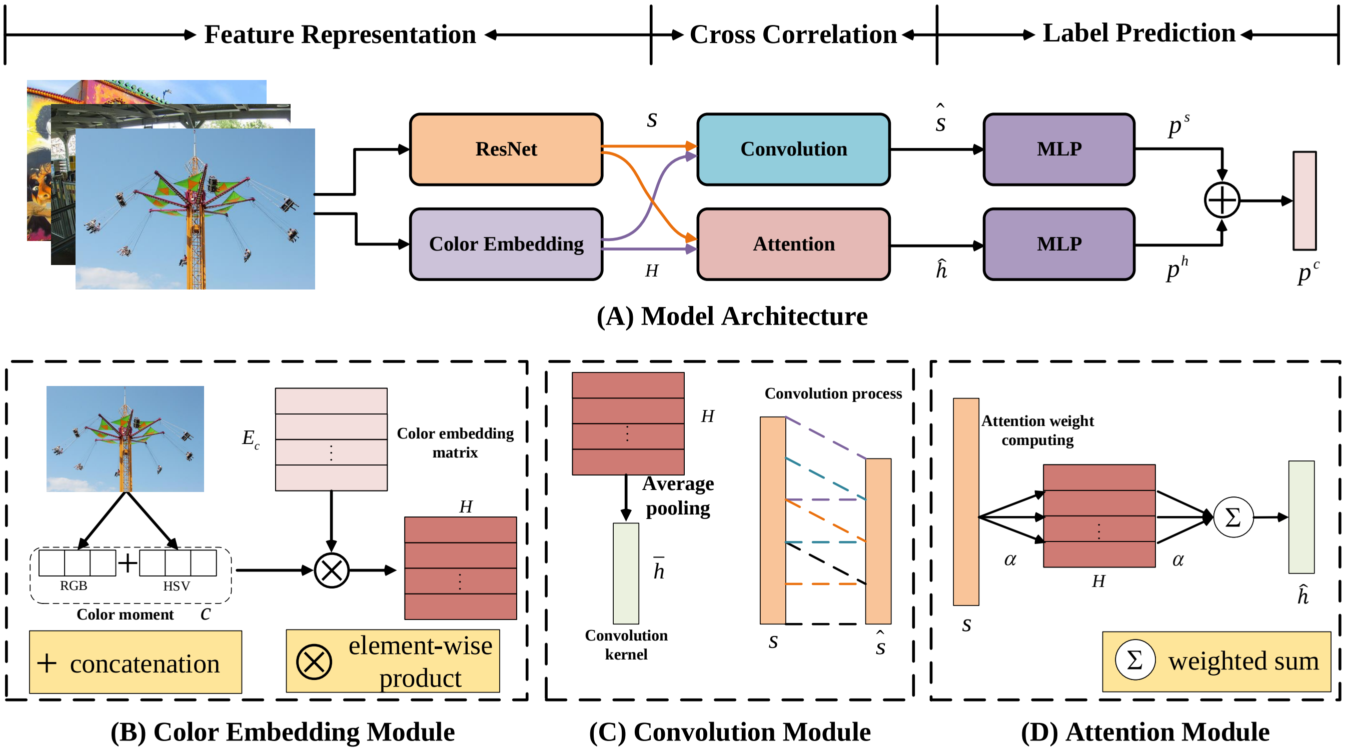


Fig. 2. Architecture of CECCN (best viewed in color). (A) The overall structure. (B) Color Embedding Module, aiming to extract color information. (C) Convolution Module, employing color features to enrich content feature representations. (D) Attention Module, utilizing content vector $\boldsymbol{s}$ to enhance color feature representations.

Content feature representation: For content features of the input image, we employ CNN to extract them since CNN is capable of recognizing the contents in the image and has achieved promising performance on many Computer Vision (CV) tasks. To be specific, in this paper, we select ResNet [6], pre-trained on ImageNet, as our content feature extractor. Since we concern more about the effects of color and content on image sentiment and their mutual influence rather than spatial factors, we use the feature vector of the last pooling layer as the image content representation rather than the matrix representations from the last convolutional layer. The extracted content feature vector $\boldsymbol{s}$ from an image $\boldsymbol{I}$ is denoted as follows:

$$\boldsymbol{s} = \mathrm{ResNet}(\boldsymbol{I}), \tag{1}$$

where $\boldsymbol{s} \in \mathbb{R}^{2048}$ means the dimension of the extracted image content feature vector is 2048.

Color feature representation: Apart from content features, color features in images also play an important role in image sentiment analysis. However, it is still full of challenges to represent color features comprehensively and integrate them with content features in the same space for subsequent calculations. Traditional statistical-based methods only map color information into a certain color space with low-dimension vector representations, which ignore the semantics of image colors that have strong relations with image sentiment. Meanwhile, CNN-based deep learning methods also have shortcomings in effective color information utilization due to the simply mixed processing of image color and content.

To tackle the above challenge, in this paper, we characterize color information with low-dimension vectors in the first step, then use an embedding method to better enrich the sentiment information represented by color and lay the foundation for calculating with content features in the same space. Color moment [42] is a simple yet effective method, which consists of first-order moment (i.e., mean), second-order moment (i.e., variance), third-order moment (i.e., skewness), and so on. Since color information is mainly distributed in the low-order moments, the first-order, second-order and third-order moments are sufficient to express the color distribution of the image. The color moments have already been proved to be effective in representing the color distribution in the image [43], [44]. Compared with histograms that cannot capture spatial relationship of color regions and have limited discriminating power [45], color moment has been shown more robust and runs faster than the histogram based methods [42] by characterizing one dimensional color distribution with the first three moments. Along this line, our proposed method could achieve better performance. To this end, we use color moment to collect the corresponding color features in the first step.

Moreover, RGB (i.e., Red, Green, Blue) color space has a wide range of applications [46], [47], and is also consistent with human intuitive perception. While HSV (i.e., Hue, Saturation, Value) space is very intuitive to express the hue, vividness and brightness of the color, which is convenient for color contrast and much closer to people's perception of color [48], [49]. To this end, we adopt color moment method to collect image color features from both RGB and HSV color space. As shown in Fig. 2 (B), this process can be formulated as follows:

$$\boldsymbol{c} = \text{ColorMoment}(\boldsymbol{I}), \tag{2}$$

where $\boldsymbol{c} \in \mathbb{R}^{18}$ since each color factor could be represented from three level moments.

As introduced before, how to integrate color information effectively into a deep sentiment framework is one of the most significant problems to be solved in this paper. Embedding method is capable of representing information with dense vectors and learning from training process, which is popular and crucial in many areas, such as NLP (Natural Language Processing) [50], Recommender System [51], Graph Representation [52] and so on. Color features extracted with color moment method are low-dimension vectors that might be weak in expressing rich sentiment semantics. To this end, we make a further step. As shown in Fig. 2 (B), we transform color feature vector $\boldsymbol{c}$ into color feature matrix $\boldsymbol{H}$ with an embedding method as follows:

$$\boldsymbol{H} = \boldsymbol{c} \otimes \boldsymbol{E_c}, \tag{3}$$

where $\boldsymbol{E_c} \in \mathbb{R}^{18\times d}$ is color embedding matrix, and d is the dimension for each color factor information. $\otimes$ means element-wise multiplication. Through the transformation, we can better enrich the sentiment information represented by color and lay the foundation for Cross Correlation in the next part.

B. Cross Correlation

In Feature Representation component, we have obtained image features from two aspects (i.e., content and color). As introduced in Section I, the image color and content factors would not only affect the sentiment expressing individually, but also influence the sentiment expressing of each other. How to leverage the correlations between these two features for better image sentiment analysis is still very challenging.

To this end, we design a novel Cross Correlation (CC) method to model the correlations of these two features for information enhancement. As shown in Fig. 2, CC consists of two branches: 1) Employing color features to enrich content feature representations with sequence convolution; 2) Utilizing content features to enhance color feature representations with attention mechanism. For the former branch, we can obtain the different content feature vectors by sequential convolution according to the color feature matrix $\boldsymbol{H}$. For the latter branch, we can obtain the attention weights of $\boldsymbol{H}_1, \boldsymbol{H}_2, ..., \boldsymbol{H}_{18}$ according to the content vector $\boldsymbol{s}$.

For the sake of clarity, we give a detailed explanation about why we adopt two different methods (i.e., sequence convolution and attention mechanism) for information enhancement when dealing with these two kinds of features (i.e., content feature and color feature) separately. To the best of our knowledge, attention mechanism generally adopts a feature vector to compute similarity with each vector in another feature matrix as attention weights, and then better represents the feature matrix as a feature vector to enhance it with the attention weights. However, when the feature information is represented with a single vector, it is of little significance to implement attention mechanism, because each dimension in the vector is a scalar and it is meaningless and improper to compute similarity between two scalars. Back to our work, since color feature is represented as a matrix in this paper, we utilize attention mechanism to obtain the attention weight distribution of color feature matrix $\boldsymbol{H}$, and then make reintegration with attention weights to get enhanced color feature vector. However, content feature extracted in Feature Representation component is represented as a vector $\boldsymbol{s}$, to which attention mechanism cannot be properly implemented. To tackle this challenge, we further develop sequence convolution to update and enhance content feature vector. We will describe the technical details in the following parts.

Image color to image content: As mentioned above, different colors will play different roles in judging the sentiment attribute of the content in the image. It seems feasible to employ color features to enrich content feature representations.

To leverage the correlations between two vectors of equal dimension for information enhancement, Tay et al. [53] defined an associative memory operator, namely circular convolution, to update the feature vector representation. Inspired by this successful work, we also design an operation called *sequence convolution*, which extends the circular convolution to deal with feature representations with different dimensions. Based on sequence convolution, CECCN could leverage the mutual influence between image color and image content to enhance image content representations for sentiment expressing. Fig. 2 (C) depicts the details of sequence convolution.

Specifically, we first transform color feature matrix $\boldsymbol{H}$ into color feature vector $\bar{\boldsymbol{h}}$ with average pooling. Then, $\bar{\boldsymbol{h}}$ is utilized as the convolution kernel for sequence convolution. The convolution process is formulated as follows:

$$\begin{aligned} \bar{\boldsymbol{h}} &= \text{avg\_pooling}(\boldsymbol{H}), \\ \hat{\boldsymbol{s}}_i &= \sum_{j=0}^{N-1} \boldsymbol{s}_{i+j}\bar{\boldsymbol{h}}_j, \end{aligned} \tag{4}$$

where $avg\_pooling(\cdot)$ means average pooling. $\bar{\boldsymbol{h}}_j$ denotes the j^{th} element of $\bar{\boldsymbol{h}}$. N is the length of vector $\bar{\boldsymbol{h}}$. $\hat{\boldsymbol{s}}$ stands for the well-learned content vector after the sequence convolution process with color vector.

Image content to image color: As mentioned in Section I, different contents may change the sentiment attributes of the same color. Thus, we take the content features into consideration when representing color information. Moreover, attention mechanism could be helpful for extracting the most relevant parts from inputs for outputs [54], [55]. Therefore, we utilize attention mechanism to obtain the attention weight distribution of color feature matrix, which can be formulated as follows:

$$\begin{aligned} \hat{\boldsymbol{h}} &= \sum_{i=1}^{18} \alpha_i \boldsymbol{H}_i, \\ \alpha_i &= \frac{exp(f(\boldsymbol{s}, \boldsymbol{H}_i))}{\sum_{j=1}^{18} exp(f(\boldsymbol{s}, \boldsymbol{H}_j))}, \\ f(\boldsymbol{s}, \boldsymbol{H}_i) &= \tanh(\boldsymbol{W}_h \boldsymbol{H}_i + \boldsymbol{W}_s \boldsymbol{s}), \end{aligned} \tag{5}$$

where α_i indicates the attention weight calculated from the content vector to the i^{th} vector of color feature matrix. $\boldsymbol{W}_h$ and $\boldsymbol{W}_s$ are trainable parameters. Subsequently, α_i is used to compute a weighted sum of the i^{th} vector of color feature matrix as $\hat{\boldsymbol{h}}$. This process is depicted in Fig. 2 (D).

C. Label Prediction

In this subsection, we make a final prediction of image sentiment for the input image. As described in the previous subsection, well-learned image representation vectors (i.e., $\hat{\boldsymbol{h}}$ and $\hat{\boldsymbol{s}}$) have been obtained with Cross Correlation.

Then, we send them to different multi-layer perceptrons (MLPs) to calculate image sentiment separately. Each MLP has two hidden layers with $Relu$ activation and a $softmax$ output layer.

$$\begin{aligned} \boldsymbol{p}^s &= \mathrm{MLP}_1(\hat{\boldsymbol{s}}), \\ \boldsymbol{p}^h &= \mathrm{MLP}_2(\hat{\boldsymbol{h}}), \end{aligned} \tag{6}$$

where $\boldsymbol{p}^s$ and $\boldsymbol{p}^h$ denote the probability distribution of color and content with respect to image sentiment separately.

In order to make the predicting result more robust, we fuse the sentiment probabilities obtained from the image content and image color to predict the final image sentiment probability $\boldsymbol{p}^c$ as follows:

$$\boldsymbol{p}^c = \lambda * \boldsymbol{p}^s + (1 - \lambda) * \boldsymbol{p}^h, \tag{7}$$

where $\lambda \in [0, 1]$ is the hyper-parameter. Its optimal value is determined by experiments.

To enhance the explanation of the model, we also summarize our proposed method in Algorithm 1.

Algorithm 1 Color Enhanced Cross Correlation Net

```
Input: Input image: $\boldsymbol{I}$
Output: Sentiment label probability distribution: $\boldsymbol{p}^c$
1: Obtain content features of $\boldsymbol{I}$ with ResNet, denote as $\boldsymbol{s}$
2: Obtain color vector of $\boldsymbol{I}$ with color moment, denote as $\boldsymbol{c}$
3: Obtain color embedding matrix $\boldsymbol{H} \in \mathbb{R}^{18 \times d}$ with embedding method
4: Employ color features $\boldsymbol{H}$ to enrich content feature representation with sequence convolution, denote as $\hat{\boldsymbol{s}}$
5: Use content features $\boldsymbol{s}$ to enhance color feature representation with attention mechanism, denote as $\hat{\boldsymbol{h}}$
6: Predict sentiment label probability distribution $\boldsymbol{p}^c$ with Cross Correlation (CC) enhanced content features $\hat{\boldsymbol{s}}$ and color features $\hat{\boldsymbol{h}}$
```

D. Model Learning

For model learning, we employ the cross-entropy as the loss function since it is a classification problem. The loss function for the output $\boldsymbol{p}^c$ of the last layer is shown as follows:

$$L^c = -\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{y}_i log P(\boldsymbol{p}_i^c | \boldsymbol{I}), \tag{8}$$

where $\boldsymbol{y}_i$ is the one-hot representation for the true class of the i^{th} instance, and N represents the number of training instances. In order to make $\boldsymbol{p}^s$, $\boldsymbol{p}^h$ also calculate the correct probability distribution, we apply cross-entropy function to both of them, the loss functions of which are denoted as L^s and L^h. Considering the model complexity, we also add L2-norm of all trainable parameters in CECCN to the final loss function, which is computed as follows:

$$L = L^s + L^h + L^c + \epsilon ||\boldsymbol{\theta}||_2, \tag{9}$$

where $\boldsymbol{\theta}$ denotes all trainable parameters in the model. Here, we also count the number of all the trainable parameters of our proposed CECCN, the total of which is only as small as 596 k and is much fewer than other state-of-the-art methods such as RA-DLNet [29] with 25 M trainable parameters.

IV. Experiment

In this section, we first introduce the experiment setup. Then, we evaluate the model performance on two public benchmark datasets for image sentiment analysis. Next, we give a detailed analysis of the model and experiment results.

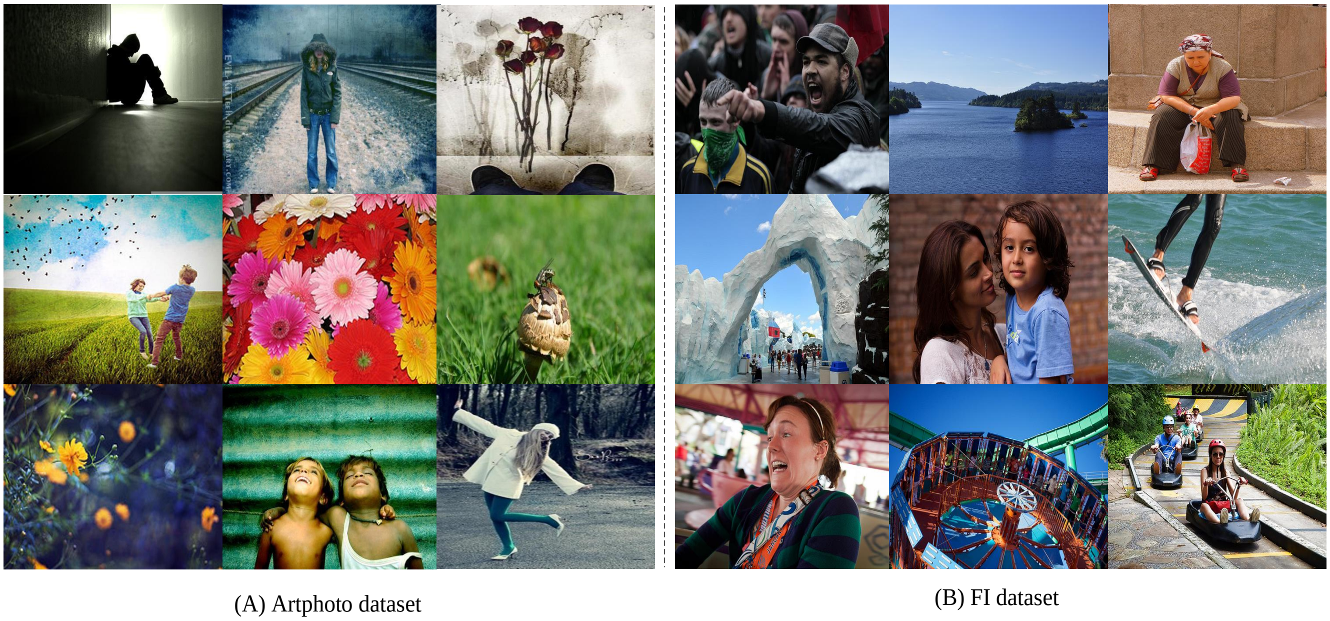
A. Datasets

As investigated, we list some current existing datasets for image sentiment analysis and their characteristics in Table I. How to choose proper data sets to validate our method is also a very important issue. In this paper, we select Artphoto and FI to verify our proposed model. The main reasons are summarized as follows:

First, considering the data style, Artphoto consists of images shared by artists who often use colors to express emotions. FI is derived from social sharing platforms, which consists of daily pictures with rich content. Second, Artphoto and FI can be selected to represent small-scale and large-scale data sets, respectively. Third, these two data sets have more fine-grained sentiment categories compared with others, which allow us to perform both coarse-grained and fine-grained classification. Therefore, we choose these two popular and well-studied benchmark data sets to verify our ideas.

- Artphoto is a publicly available dataset with 806 artistic photos and eight sentiment categories defined by [60]. It is obtained by using the eight emotion categories as search terms in the art sharing site[1]. These photos are taken by people who attempt to evoke a certain emotion in the viewer of the photo through the conscious manipulation of the image lighting, colors, etc. This dataset therefore allows us to investigate whether the conscious use of colors by artists improves the classification.
- FI has 23,308 images collected by You et al. [58] They queried images with those eight sentiment categories as keywords from Flickr and Instagram. In this way, weakly labeled images are collected. Next, they deleted images which have tags of any two different emotions. Then, they employed Amazon Mechanical Turk (AMT) to further label these weakly labeled images which result in 23,308 images receiving at least three agreements. Since images in FI are all copyrighted by laws, some links or images

[1] https://www.deviantart.com/


(A) Artphoto dataset

(B) FI dataset

Fig. 3. Example images from Artphoto and FI datasets. The images come from a variety of domains. Among them, Artphoto contains more images in art domain, while images in FI are mostly taken from real life.

TABLE I
STATISTICS OF SOME CURRENT EXISTED DATASETS FOR IMAGE SENTIMENT ANALYSIS.

Dataset	Size	Label	Source
IAPS-Subset [56]	395	awe, amusement, contentment, excitement, disgust, anger, fear, sad	International Affective Picture System
Abstract Paintings [14]	228	awe, amusement, contentment, excitement, disgust, anger, fear, sad	peer rated abstract paintings
Artphoto [14]	806	awe, amusement, contentment, excitement, disgust, anger, fear, sad	DeviantArt, uploaded by artists
EmotionROI [57]	1980	anger, disgust, joy, fear, sadness, surprise	Flickr
FI [58]	23,308	awe, amusement, contentment, excitement, disgust, anger, fear, sad	Flickr and Instagram
CrossSentiment [32]	155,578	positive, negative, neutral	Flickr and Instagram
T4SA [59]	1.5 M	positive, negative, neutral	Twitter

TABLE II
STATISTICS OF THE ARTPHOTO AND FI DATASETS ON BOTH TWO-CLASS CATEGORIES AND EIGHT-CLASS CATEGORIES.

Two-class categories	**Eight-class categories**	**Dataset**	
		Artphoto	**FI**
positive	amusement	101	4724
	awe	102	2881
	contentment	70	5129
	excitement	105	2725
negative	anger	77	1176
	disgust	70	1591
	fear	115	969
	sadness	166	2633
Total Count		806	21828

are corrupted now. In this paper, we remove these samples and retain 21,828 images.

For both two datasets, they are randomly split into 80% training and 20% testing set. As an emphasis, we conduct experiments on both eight-class and two-class sentiment classification tasks. The former one directly uses labels in datasets, while the latter one divides eight sentiment categories into binary labels according to [56], [27], which suggests that *amusement*, *awe*, *contentment* and *excitement* are *positive* sentiments, while *anger*, *disgust*, *fear* and *sadness* are *negative* sentiments. In Table II, we give detailed statistics of the Artphoto and FI datasets on data distribution. In Fig. 3, we list some example images from these two datesets.

B. Implementation Details

- **Model Setting**: In our work, for image content feature extraction, we employ ResNet [6] pre-trained on large scale object classification dataset ImageNet [7], as basic CNN architecture. Images are fed into ResNet with the resolution of 448×448. For image color feature extraction, we empirically set the dimension (i.e., d) of color embedding as 100.
- **Training Setting**: To initialize the model, we randomly set all weights such as $\boldsymbol{W}$ following the truncated normal distribution, where mean equals to 0 and standard deviation is set to 0.1. We use Adam optimizer with the learning rate of 5×10^{-4}. During the implementation, we utilize *Tensorflow* to build our entire model.

TABLE III
OVERALL PERFORMANCE (ACCURACY) OF DIFFERENT MODELS.

Model	**Artphoto**		**FI**	
	Two-class Test	**Eight-class Test**	**Two-class Test**	**Eight-class Test**
(1) GCH [16]	56.44%	21.78%	70.83%	27.05%
(2) PCNN [25]	68.81%	31.68%	75.34%	46.09%
(3) SentiNet-A [9]	72.35%	35.40%	78.74%	46.87%
(4) ResNet [6]	73.76%	33.17%	83.56%	53.29%
(5) AR [27]	74.80%	32.67%	86.35%	53.72%
(6) $GM_{EI}\&LRM_{SI}$ [10]	72.86%	35.15%	88.02%	53.47%
(7) **RA-DLNet** (ResNet version) [29]	80.86%	34.01%	87.01%	56.18%
(8) WSCNet [28]	80.25%	30.25%	88.25%	**68.42%**
(9) CECCN	**83.33%**	**43.83%**	**88.55%**	67.96%

- **Evaluation Metrics**: Following [9], [27], [10], for better comparison, we also adopt *accuracy* as our main evaluation metric for both two-class and eight-class classification tasks on two datasets. Moreover, as [29] did, considering the unbalanced nature of Artphoto, we also use True Positive Rate (TPR) to demonstrate the performance on Artphoto. For eight-class test, we additionally adopt Confusion Matrix to show the results.

C. Baselines

In this paper, we compare our model against the following state-of-the-art baselines:

- **GCH** [16]: Extracting global image color information with color histogram.
- **PCNN** [25]: Leveraging a progressive training strategy and a domain transfer strategy to fine-tune the pre-trained CNN for sentiment classification.
- **SentiNet-A** [9]: Integrating visual attention into the CNN, and employing saliency map as a prior knowledge and regularizer to holistically refine the attention distribution for sentiment prediction.
- **ResNet** [6]: Simply utilizing ResNet, which is proven to be strong in image content classification, to obtain image features as a baseline for image sentiment analysis.
- **AR** [27]: Automatically discovering effective regions by taking the objectness score as well as the sentiment score into consideration, and aggregating CNN outputs from local regions and the whole images to produce the final sentiment prediction.
- $\boldsymbol{GM_{EI}\&LRM_{SI}}$ [10]: Using both global image and local information of salient objects, with two models trained independently to handle the corresponding images.
- **WSCNet** [28]: Detecting a sentiment specific soft map by training a fully convolutional network with the cross spatial pooling strategy in the detection branch. And both the holistic and localized information are utilized by coupling the sentiment map with deep features as semantic vector.
- **RA-DLNet** [29]: Firstly utilizing a pre-trained CNN (e.g., VGG [61], ResNet [6], InceptionNet [62], NasNet [63]) to extract image features, and then applying residual attention model to focus on crucial sentiment-rich, local regions in the image. Note that we choose the ResNet version RA-DLNet as baseline for an fair comparison.

Among these baselines, GCH is a color based method, while the others are all content based models. Specially, $GM_{EI}\&LRM_{SI}$, RA-DLNet and WSCNet are the current state-of-the-art models in image sentiment analysis.

D. Experiment Results

In this subsection, we evaluate our model on Artphoto and FI datasets with both two-class and eight-class sentiment classification tasks. For the sake of fairness, we utilize *Tensorflow* to implement the baseline models which are not yet public available, and utilize the published code if the model source code is available. Under the dataset split in this paper, we present the optimal results in Table III.

As illustrated in Table III, CECCN achieves highly comparable performance on all test sets. Specifically, CECCN takes both image content and image color into consideration, since they are able to provide complementary information for sentiment analysis. Moreover, CC is designed for information enhancement to better analyze image sentiment, by leveraging the correlations between these two different features. In particular, CECCN has a significant performance improvement on small scale data set (e.g., Artphoto) and fine-grained classification task (i.e., eight-class test) compared to baselines.

Among these baselines, GCH is an earlier traditional method that simply utilizes color histogram to extract color features. It overlooks content information of an image, which leads to poor performance on test sets. PCNN and SentiNet-A attempt to apply deep learning methods by employing pre-trained CNNs to sentiment analysis and have achieved great success against traditional methods. ResNet seeks a simple way to extract image content features as a vector and make sentiment classification with MLP.

Aiming at capturing more image content details, large efforts have been made to leverage both global and local information of an image in AR, $GM_{EI}\&LRM_{SI}$ and WSCNet. The outstanding performance proves the importance of image content for sentiment analysis. RA-DLNet further adopt residual attention to enhance the image understanding. However, they both rely on pre-trained CNNs to extract content information and fail to make full use of color information, which leads to limited performance, especially on small scale data set (e.g., Artphoto) and fine-grained classification task (i.e., eight-class classification).

From Table III, we could obtain another interesting phenomenon. For fine-grained classification, some state-of-the-art

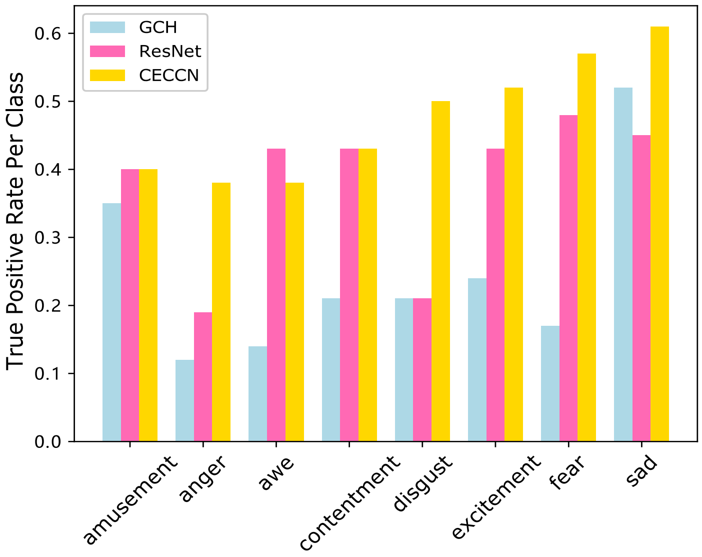
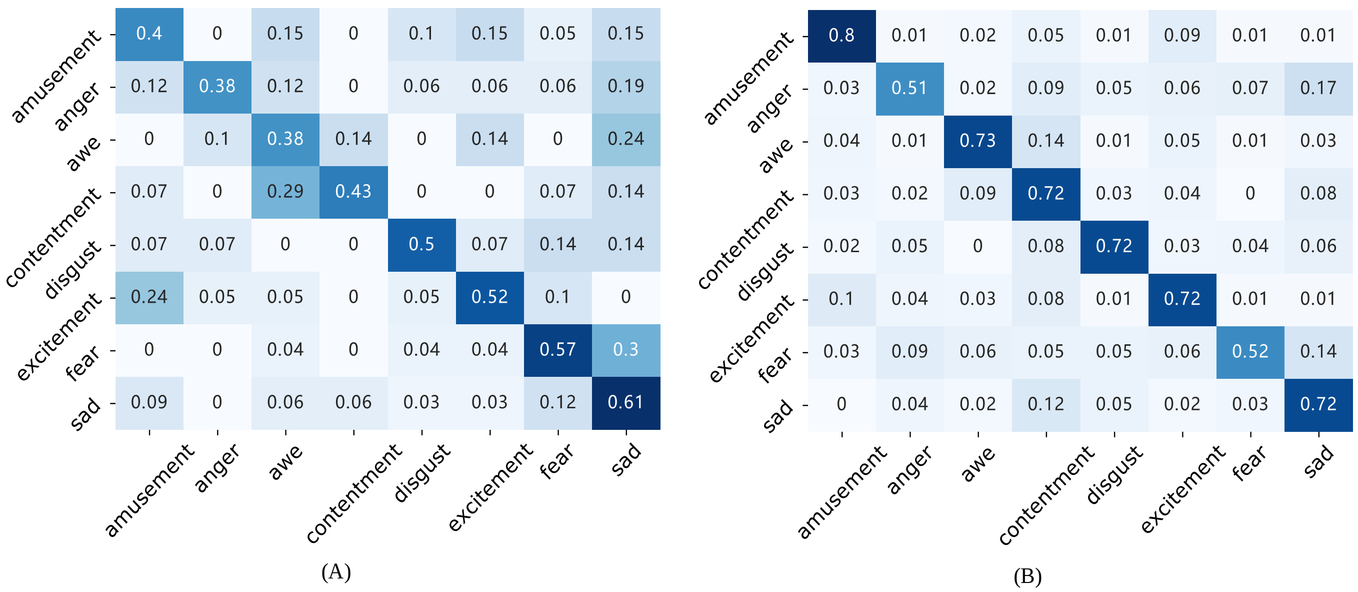


Fig. 4. True Positive Rate (TPR) of GCH [16], ResNet [6] and our proposed CECCN on Artphoto dataset.



	amusement	anger	awe	contentment	disgust	excitement	fear	sad
amusement	0.4	0	0.15	0	0.1	0.15	0.05	0.15
anger	0.12	0.38	0.12	0	0.06	0.06	0.06	0.19
awe	0	0.1	0.38	0.14	0	0.14	0	0.24
contentment	0.07	0	0.29	0.43	0	0	0.07	0.14
disgust	0.07	0.07	0	0	0.5	0.07	0.14	0.14
excitement	0.24	0.05	0.05	0	0.05	0.52	0.1	0
fear	0	0	0.04	0	0.04	0.04	0.57	0.3
sad	0.09	0	0.06	0.06	0.03	0.03	0.12	0.61

(A)

	amusement	anger	awe	contentment	disgust	excitement	fear	sad
amusement	0.8	0.01	0.02	0.05	0.01	0.09	0.01	0.01
anger	0.03	0.51	0.02	0.09	0.05	0.06	0.07	0.17
awe	0.04	0.01	0.73	0.14	0.01	0.05	0.01	0.03
contentment	0.03	0.02	0.09	0.72	0.03	0.04	0	0.08
disgust	0.02	0.05	0	0.08	0.72	0.03	0.04	0.06
excitement	0.1	0.04	0.03	0.08	0.01	0.72	0.01	0.01
fear	0.03	0.09	0.06	0.05	0.05	0.06	0.52	0.14
sad	0	0.04	0.02	0.12	0.05	0.02	0.03	0.72

(B)

Fig. 5. Confusion matrix for Artphoto (A) and FI (B).

methods could achieve very good performance in large-scale data set (e.g., FI) which might benefit a lot from very deep models. However, in small-scale data set (e.g., Artphoto), these methods might not perform well compared with some earlier models (e.g., SentiNet-A and ResNet). Therefore, we speculate that those methods might depend on the data set scale when dealing with fine-grained classification task, and our proposed CECCN could be robust against the variance of data set scale. Besides, Artphoto is a data set full of artistic flavor and style. The importance of color to image sentiment expressing is more obvious in this data set. And CECCN especially considers and explores the color and the relationship between color and content regarding the interaction and collaboration of image sentiment. Therefore, CECCN could achieve much better performance on Artphoto dataset compared with other baseline methods.

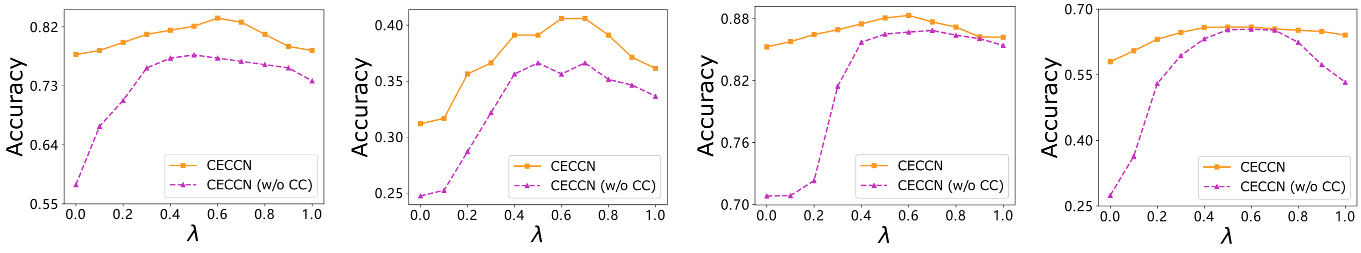
In addition, the TPR performance comparisons on Artphoto dataset are depicted in Fig. 4 considering its unbalanced nature. We also summarize the eight-class test results with confusion matrix in Fig. 5. These results again demonstrate the effectiveness of our proposed CECCN.

E. Ablation Performance

The overall experimental results have already proven the superiority of our proposed CECCN method. However, which component is really important for performance improvement is still unclear. Thus, in this subsection, we conduct an ablation study on CECCN to examine the effectiveness of each component. The results are illustrated in Table IV.

TABLE IV
ABLATION STUDY OF CECCN, WHERE W/O MEANS WITHOUT.

Model	Artphoto		FI	
	Two-class Test	Eight-class Test	Two-class Test	Eight-class Test
(1) CECCN (w/o color)	73.76%	33.17%	83.56%	53.29%
(2) CECCN (w/o content)	57.92%	24.75%	70.83%	26.13%
(3) CECCN (w/o CC)	78.22%	37.62%	86.20%	64.19%
(4) CECCN (w/o CC)+MLP	76.73%	34.65%	86.70%	64.27%
(5) CECCN	**83.33%**	**43.83%**	**88.55%**	**67.96%**



(A) 2-class Test on Artphoto (B) 8-class Test on Artphoto (C) 2-class Test on FI (D) 8-class Test on FI

Fig. 6. Parameter sensitivity study of CECCN on different λ settings, where w/o means without.

As mentioned before, we utilize content features and color features simultaneously for complementary information. The former one is one of the essential parts in plenty of CV applications, while the latter one could report the approximate sentiment polarity. As shown in Table IV (1)-(2), the performance of CECCN significantly decreased when removing them separately, which means both image color and image content are critical for image sentiment analysis.

Recalling the model architecture, the color features and content features are cross correlated in the model after extracting them separately. We are curious whether CC is really crucial for CECCN. To this end, we design two validation experiments. In the first step, we remove CC module to verify it. The results in Table IV (3) illustrate that CC is capable of building up correlations between these content features and color features for information enhancement, which is very important for image sentiment analysis. Then, we further replace CC with two MLPs (i.e., one for attention mechanism, the other for sequence convolution). Experiment results in Table IV (4) demonstrate that CC is indispensable for CECCN to achieve better performance in image sentiment analysis, especially for small scale dataset (e.g., Artphoto).

F. Sensitivity Analysis and Robustness Test

As mentioned before, the hyper-parameter λ in Label Prediction component controls the importance of content features and color features for the final decision. We intend to figure out how this parameter affects model performance. Thus, we conduct parameter sensitivity experiments in this subsection. Fig. 6 illustrates the corresponding results.

From the experimental results, we observe that for different classification tasks and datasets, the hyper-parameter λ has different effects on the performance of CECCN. Overall, the performance of CECCN first becomes better with the increase of λ. When λ is between 0.5 and 0.7, CECCN achieves the best performance. When λ is bigger than 0.7 or so, the accuracy decreases to varying degrees. This phenomenon is consistent with the observations in Section I that both image content and image color are important for image sentiment analysis, and they are able to provide complementary information. If the proportion of either content feature or color feature is too low, model performance will degrade. Thus, they should be effectively integrated for better image sentiment analysis.

Moreover, we conduct another set of experiments, in which CC is removed from the entire model. The purple dashed lines in Fig. 6 show the results. From the results, we can observe that the overall trend of experimental performance with the change of λ is consistent with the full model, but the change is a bit greater. Compared with the performance of CECCN, we could draw the conclusion that CC is also capable of improving the robustness of the model and reducing the impact of hyper-parameter tuning on model performance.

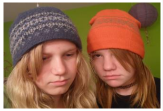
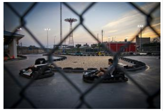
G. Case Study and Error Analysis

In order to better evaluate the model performance, in this subsection, we select some examples from test sets to demonstrate the ability of CECCN to leverage both image content and color information comprehensively. The results are presented in Fig. 7.

Taking the first image in Fig. 7 (A) as an example, if only color information is considered without content information, the model misclassifies the sentiment as *Negative*. We speculate the reason is that the extracted color information is a relatively dark hue, consequently, the model simply predicted negative sentiment. However, when taking content into consideration, *Positive* would be correctly predicted, since beautiful fireworks usually demonstrate a positive sentiment. For the second image in Fig. 7 (A), a boy was swinging, which might lead to a *Positive* sentiment if only image content was taken into account. However, when we consider color information as a very important factor, we will get the correct classification, which might be because the dark night

Image	w/o Color	w/o Content	CECCN	Gold Label
	Positive	Negative	Positive	Positive
	Positive	Negative	Negative	Negative

(A)

Image	w/o Att	w/o Cov	CECCN	Gold Label
	Positive	Positive	Negative	Negative
	Negative	Positive	Positive	Positive

(B)

Fig. 7. Examples of classification results for different models. In (B), Att means attention mechanism, and Cov stands for sequence convolution in CC. For the sake of better comparison, we mark the misclassified labels with red font.

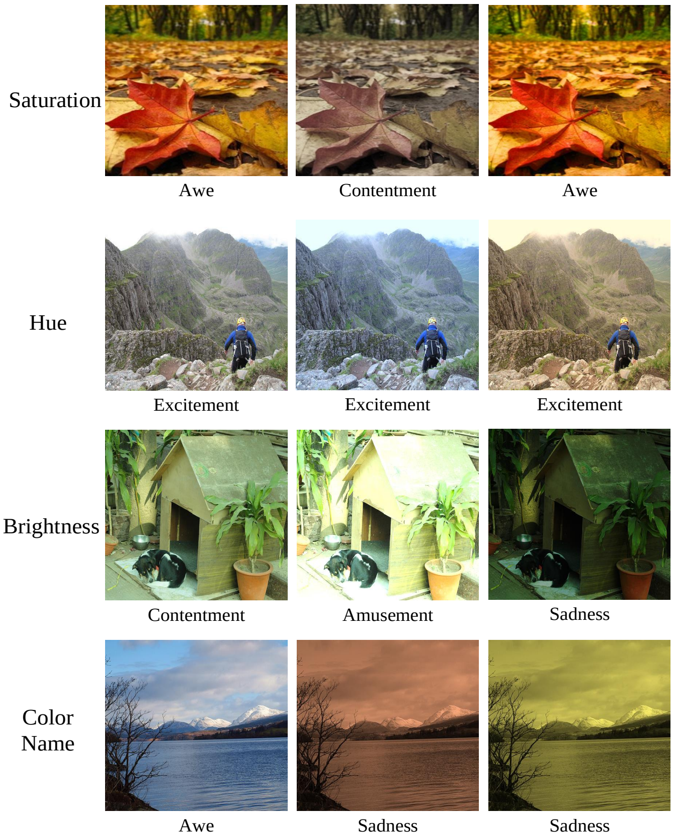


Fig. 8. Cases for randomly changing the relative image color factors.

indicating *Negative* provides important information for image sentiment classification.

Experiment results in Fig. 7 (B) show that removing either attention module or convolution module in CC might lead to incorrect predictions, since they overlook the correlation and mutual influence between content information and color information. As we have observed before, different colors will change the sentiment attributes of the same content, so does the content. We speculate that the coordination and correlation of content and color are also important for image sentiment. By using CC, we could better integrate the two kinds of learned features and make better image sentiment analysis.

As we extract color features from multiple aspects and have proved that image color is important for sentiment analysis in the previous parts, we are still curious about how colors change image sentiment and how CECCN will respond to the changes. To this end, we conduct further exploration concerning color factors. We randomly change the relative image color factors and demonstrate how the model will react to the corresponding change in Fig. 8.

For a succinct analysis, take an example in the first row of Fig. 8. When *saturation* changes, CECCN is capable of perceiving the change and the results are also quite in line with human perception. The effect of *brightness* and *color name* changes is similar. Taking a special look in the second row of Fig. 8. As the *hue* of the image varies, the sentiment might not necessarily change accordingly. This is also a reasonable phenomenon, since CECCN will comprehensively consider both the color and content, including the interaction and coordination between them.

V. Conclusion

In this paper, we argue that image color is a critical factor for image sentiment analysis, which has not been analyzed comprehensively. To this end, we proposed a Color Enhanced Cross Correlation Net (CECCN) for image sentiment analysis, in which both image content and color information were taken into account to provide complementary information. Along this line, inspired by the observation that contents and colors might affect the sentiment expressing of each other, we further developed a Cross Correlation (CC) method to model the correlations between contents and colors for information enhancement, with attention mechanism and sequence convolution. Extensive experimental results on two publicly available datasets and two different classification tasks demonstrated the superiority and rationality of CECCN.

Along this line, we will explore how to make use of color effects on image sentiment more comprehensively. In the future, we will also conduct deeper research with a special focus on the implicit factors that influence image sentiment analysis, such as object relations and label relations.

Acknowledgment

This research was partially supported by grants from the National Natural Science Foundation of China (Grants No. 61727809, 61922073, 62006066) and the Open Project Program of the National Laboratory of Pattern Recognition (NLPR).

REFERENCES

[1] 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, San Francisco, CA, USA, February 22-25, 2016*, P. N. Bennett, V. Josifovski, J. Neville, and F. Radlinski, Eds. ACM, 2016, pp. 13–22. [Online]. Available: https://doi.org/10.1145/2835776.2835779

[2] L. Pang, S. Zhu, and C.-W. Ngo, "Deep multimodal learning for affective analysis and retrieval," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2008–2020, 2015.

[3] D. Hyun, C. Park, M. Yang, I. Song, J. Lee, and H. Yu, "Review sentiment-guided scalable deep recommender system," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, K. Collins-Thompson, Q. Mei, B. D. Davison, Y. Liu, and E. Yilmaz, Eds. ACM, 2018, pp. 965–968. [Online]. Available: https://doi.org/10.1145/3209978.3210111

[4] S. Zhao, Y. Gao, G. Ding, and T. Chua, "Real-time multimedia social event detection in microblog," *IEEE Trans. Cybernetics*, vol. 48, no. 11, pp. 3218–3231, 2018. [Online]. Available: https://doi.org/10.1109/TCYB.2017.2762344

[5] B. Pang, L. Lee *et al.*, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[7] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, 2009, pp. 248–255. [Online]. Available: https://doi.org/10.1109/CVPR.2009.5206848

[8] V. Campos, B. Jou, and X. Giro-i Nieto, "From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction," *Image and Vision Computing*, vol. 65, pp. 15–22, 2017.

[9] K. Song, T. Yao, Q. Ling, and T. Mei, "Boosting image sentiment analysis with visual attention," *Neurocomputing*, vol. 312, pp. 218–228, 2018.

[10] L. Wu, M. Qi, M. Jian, and H. Zhang, "Visual sentiment analysis by combining global and local information," *Neural Processing Letters*, pp. 1–13, 2019.

[11] K. W. Jacobs and F. E. Hustmyer Jr, "Effects of four psychological primary colors on gsr, heart rate and respiration rate," *Perceptual and motor skills*, vol. 38, no. 3, pp. 763–766, 1974.

[12] G. D. Wilson, "Arousal properties of red versus green." *Perceptual and motor skills*, 1966.

[13] J. P. Guilford and P. C. Smith, "A system of color-preferences," *The American Journal of Psychology*, vol. 72, no. 4, pp. 487–502, 1959.

[14] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, A. D. Bimbo, S. Chang, and A. W. M. Smeulders, Eds. ACM, 2010, pp. 83–92. [Online]. Available: https://doi.org/10.1145/1873951.1873965

[15] K. W. Jacobs and J. F. Suess, "Effects of four psychological primary colors on anxiety state," *Perceptual and motor skills*, vol. 41, no. 1, pp. 207–210, 1975.

[16] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, "Analyzing and predicting sentiment of images on the social web," in *MM*. ACM, 2010, pp. 715–718.

[17] X. Lu, P. Suryanarayan, R. B. A. Jr., J. Li, M. G. Newman, and J. Z. Wang, "On shape and the computability of emotions," in *Proceedings of the 20th ACM Multimedia Conference, MM '12, Nara, Japan, October 29 - November 02, 2012*, N. Babaguchi, K. Aizawa, J. R. Smith, S. Satoh, T. Plagemann, X. Hua, and R. Yan, Eds. ACM, 2012, pp. 229–238. [Online]. Available: https://doi.org/10.1145/2393347.2393384

[18] S. Zhao, Y. Gao, X. Jiang, H. Yao, T. Chua, and X. Sun, "Exploring principles-of-art features for image emotion recognition," in *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*, K. A. Hua, Y. Rui, R. Steinmetz, A. Hanjalic, A. Natsev, and W. Zhu, Eds. ACM, 2014, pp. 47–56. [Online]. Available: https://doi.org/10.1145/2647868.2654930

[19] D. Borth, R. Ji, T. Chen, T. M. Breuel, and S. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *ACM Multimedia Conference, MM '13, Barcelona, Spain, October 21-25, 2013*. ACM, 2013, pp. 223–232. [Online]. Available: https://doi.org/10.1145/2502081.2502282

[20] Y. Song, S. Chen, Y. Zhao, and Q. Jin, "Unpaired cross-lingual image caption generation with self-supervised rewards," in *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, L. Amsaleg, B. Huet, M. A. Larson, G. Gravier, H. Hung, C. Ngo, and W. T. Ooi, Eds. ACM, 2019, pp. 784–792. [Online]. Available: https://doi.org/10.1145/3343031.3350996

[21] L. Guo, J. Liu, J. Tang, J. Li, W. Luo, and H. Lu, "Aligning linguistic words and visual semantic units for image captioning," in *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, L. Amsaleg, B. Huet, M. A. Larson, G. Gravier, H. Hung, C. Ngo, and W. T. Ooi, Eds. ACM, 2019, pp. 765–773. [Online]. Available: https://doi.org/10.1145/3343031.3350943

[22] F. Liu, J. Liu, R. Hong, and H. Lu, "Erasing-based attention learning for visual question answering," in *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, L. Amsaleg, B. Huet, M. A. Larson, G. Gravier, H. Hung, C. Ngo, and W. T. Ooi, Eds. ACM, 2019, pp. 1175–1183. [Online]. Available: https://doi.org/10.1145/3343031.3350993

[23] F. Liu, J. Liu, Z. Fang, R. Hong, and H. Lu, "Densely connected attention flow for visual question answering," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, S. Kraus, Ed. ijcai.org, 2019, pp. 869–875. [Online]. Available: https://doi.org/10.24963/ijcai.2019/122

[24] A. Ortis, G. M. Farinella, and S. Battiato, "Survey on visual sentiment analysis," *IET Image Processing*, vol. 14, no. 8, pp. 1440–1456, 2020.

[25] 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, January 25-30, 2015, Austin, Texas, USA*, B. Bonet and S. Koenig, Eds. AAAI Press, 2015, pp. 381–388. [Online]. Available: http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9556

[26] W. Zhang, X. He, and W. Lu, "Exploring discriminative representations for image emotion recognition with cnns," *IEEE Transactions on Multimedia*, 2019.

[27] J. Yang, D. She, M. Sun, M.-M. Cheng, P. L. Rosin, and L. Wang, "Visual sentiment prediction based on automatic discovery of affective regions," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2513–2525, 2018.

[28] D. She, J. Yang, M.-M. Cheng, Y.-K. Lai, P. L. Rosin, and L. Wang, "Wscnet: Weakly supervised coupled networks for visual sentiment classification and detection," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1358–1371, 2019.

[29] A. Yadav and D. K. Vishwakarma, "A deep learning architecture of ra-dlnet for visual sentiment analysis," *Multimedia Systems*, vol. 26, pp. 431–451, 2020.

[30] R. Zheng, W. Li, and Y. Wang, "Visual sentiment analysis by leveraging local regions and human faces," in *International Conference on Multimedia Modeling*. Springer, 2020, pp. 303–314.

[31] J. Zhang, M. Chen, H. Sun, D. Li, and Z. Wang, "Object semantics sentiment correlation analysis enhanced image sentiment classification," *Knowledge-Based Systems*, vol. 191, p. 105245, 2020.

[32] M. Katsurai and S. Satoh, "Image sentiment analysis using latent correlations among visual, textual, and sentiment views," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2837–2841.

[33] A. Felicetti, M. Martini, M. Paolanti, R. Pierdicca, E. Frontoni, and P. Zingaretti, "Visual and textual sentiment analysis of daily news social media images by deep learning," in *International Conference on Image Analysis and Processing*. Springer, 2019, pp. 477–487.

[34] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *International journal of computer vision*, vol. 116, no. 1, pp. 1–20, 2016.

[35] X. Zhu, B. Cao, S. Xu, B. Liu, and J. Cao, "Joint visual-textual sentiment analysis based on cross-modality attention mechanism," in *International conference on multimedia modeling*. Springer, 2019, pp. 264–276.

[36] A. Ortis, G. M. Farinella, G. Torrisi, and S. Battiato, "Exploiting objective text description of images for visual sentiment analysis," *Multimedia Tools and Applications*, pp. 1–24, 2020.

[37] F. Huang, X. Zhang, Z. Zhao, J. Xu, and Z. Li, "Image–text sentiment analysis via deep multimodal attentive fusion," *Knowledge-Based Systems*, vol. 167, pp. 26–37, 2019.

[38] M. P. Fortin and B. Chaib-Draa, "Multimodal sentiment analysis: A multitask learning approach." in *ICPRAM*, 2019, pp. 368–376.

[39] S. Corchs, E. Fersini, and F. Gasparini, "Ensemble learning on visual and textual data for social image emotion classification," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 8, pp. 2057–2070, 2019.

[40] P. J. Profusek and D. W. Rainey, "Effects of baker-miller pink and red on state anxiety, grip strength, and motor precision," *Perceptual and motor skills*, vol. 65, no. 3, pp. 941–942, 1987.

[41] P. Valdez and A. Mehrabian, "Effects of color on emotions." *Journal of experimental psychology: General*, vol. 123, no. 4, p. 394, 1994.

[42] M. A. Stricker and M. Orengo, "Similarity of color images," in *Storage and retrieval for image and video databases III*, vol. 2420. International Society for Optics and Photonics, 1995, pp. 381–392.

[43] Z.-C. Huang, P. P. Chan, W. W. Ng, and D. S. Yeung, "Content-based image retrieval using color moment and gabor texture feature," in *2010 International Conference on Machine Learning and Cybernetics*, vol. 2. IEEE, 2010, pp. 719–724.

[44] M. Gong, Y. Hao, H. Mo, and H. Li, "Naturally combined shape-color moment invariants under affine transformations," *Computer Vision and Image Understanding*, vol. 162, pp. 46–56, 2017.

[45] S. M. Singh and K. Hemachandran, "Image retrieval based on the combination of color histogram and color moment," *International Journal of Computer Applications*, vol. 58, no. 3, 2012.

[46] P. Liu, J.-M. Guo, K. Chamnongthai, and H. Prasetyo, "Fusion of color histogram and lbp-based features for texture image retrieval and classification," *Information Sciences*, vol. 390, pp. 95–111, 2017.

[47] F. Nie, J. Li, and X. Li, "Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, S. Kambhampati, Ed. IJCAI/AAAI Press, 2016, pp. 1881–1887. [Online]. Available: http://www.ijcai.org/Abstract/16/269

[48] T. Zhang, H.-M. Hu, and B. Li, "A naturalness preserved fast dehazing algorithm using hsv color space," *IEEE Access*, vol. 6, pp. 10 644–10 649, 2018.

[49] O. R. Indriani, E. J. Kusuma, C. A. Sari, E. H. Rachmawanto *et al.*, "Tomatoes classification using k-nn based on glcm and hsv color space," in *2017 international conference on innovative and creative information technology (ICITech)*. IEEE, 2017, pp. 1–6.

[50] S. Cascianelli, G. Costante, A. Devo, T. A. Ciarfuglia, P. Valigi, and M. L. Fravolini, "The role of the input in natural language video description," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 271–283, 2019.

[51] X. Du, X. He, F. Yuan, J. Tang, Z. Qin, and T.-S. Chua, "Modeling embedding dimension correlations via convolutional neural collaborative filtering," *ACM Transactions on Information Systems (TOIS)*, vol. 37, no. 4, pp. 1–22, 2019.

[52] H. Cai, V. W. Zheng, and K. C.-C. Chang, "A comprehensive survey of graph embedding: Problems, techniques, and applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 9, pp. 1616–1637, 2018.

[53] Y. Tay, L. A. Tuan, and S. C. Hui, "Learning to attend via word-aspect associative fusion for aspect-based sentiment analysis," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, S. A. McIlraith and K. Q. Weinberger, Eds. AAAI Press, 2018, pp. 5956–5963. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16570

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[55] L. Peng, Y. Yang, Z. Wang, X. Wu, and Z. Huang, "Cra-net: Composed relation attention network for visual question answering," in *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, L. Amsaleg, B. Huet, M. A. Larson, G. Gravier, H. Hung, C. Ngo, and W. T. Ooi, Eds. ACM, 2019, pp. 1202–1210. [Online]. Available: https://doi.org/10.1145/3343031.3350925

[56] 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*, vol. 37, no. 4, pp. 626–630, 2005.

[57] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen, "Where do emotions come from? predicting the emotion stimuli map," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 614–618.

[58] 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*, vol. 30, no. 1, 2016.

[59] L. Vadicamo, F. Carrara, A. Cimino, S. Cresci, F. Dell'Orletta, F. Falchi, and M. Tesconi, "Cross-media learning for image sentiment analysis in the wild," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 308–317.

[60] 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*. IEEE, 2008, pp. 101–104.

[61] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[62] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[63] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697–8710.

Shulan Ruan received the B.S. degree from Hunan University, Changsha, China, in 2018. He is currently working toward the Ph.D. degree with the School of Computer Science and Technology, University of Science and Technology of China, Hefei, China. His research interests include sentiment analysis, computer vision and natural language processing. He has published several papers in AAAI and ICME.

Kun Zhang received the Ph.D. degree in computer science and technology from University of Science and Technology of China, Hefei, China, in 2019. He is is currently a faculty member with the Hefei University of Technology (HFUT), China. His research interests include natural language processing, Recommendation System, and text mining. He has published several papers in refereed conference proceedings such as AAAI, KDD, ICDM. He received the KDD 2018 Best Student Paper Award.

Le Wu received the Ph.D. degree in computer science from the University of Science and Technology of China (USTC). She is an associate professor with the Hefei University of Technology (HFUT), China. Her general area of research is data mining, recommender system, and social network analysis. She has published several papers in referred journals and conferences, such as the IEEE Transactions on Knowledge and Data Engineering, the ACM Transactions on Intelligent Systems and Technology, AAAI, IJCAI, KDD, SDM, and ICDM. She is the recipient of the Best of SDM 2015 Award.

Tong Xu (M'17) received the Ph.D. degree in University of Science and Technology of China (USTC), Hefei, China, in 2016. He is currently working as an Associate Professor of the Anhui Province Key Laboratory of Big Data Analysis and Application, USTC. He has authored 60+ journal and conference papers in the fields of social network and social media analysis, including IEEE TKDE, IEEE TMC, IEEE TMM, KDD, AAAI, ICDM, etc.

Qi Liu received the Ph.D. degree in computer science from USTC. He is a professor with USTC. His general area of research is datamining and knowledge discovery. He has published prolifically in refereed journals and conference proceedings, e.g., the IEEE Transactions on Knowledge and Data Engineering, the ACM Transactions on Information Systems, the ACM Transactions on Knowledge Discovery from Data, the ACM Transactions on Intelligent Systems and Technology, KDD, IJCAI, AAAI, ICDM, SDM, and CIKM. He is a member of the ACM and the IEEE. He received the ICDM 2011 Best Research Paper Award and the Best of SDM 2015 Award.

Enhong Chen (SM'07) received the Ph.D. degree from USTC. He is a professor and vice dean of the School of Computer Science, USTC. His general area of research includes data mining and machine learning, social network analysis, and recommender systems. He has published more than 100 papers in refereed conferences and journals, including the IEEE Transactions on Knowledge and Data Engineering, the IEEE Transactions on Mobile Computing, KDD, ICDM, NIPS, and CIKM. He was on program committees of numerous conferences including KDD, ICDM, and SDM. His research is supported by the National Science Foundation for Distinguished Young Scholars of China. He is a senior member of the IEEE.