Adaptive Deep Metric Learning for Affective Image Retrieval and Classification

Xingxu Yao*, Dongyu She*, Haiwei Zhang, Jufeng Yang, Ming-Ming Cheng and Liang Wang

***Abstract*—An image is worth a thousand words. Many researchers have conducted extensive studies to understand visual emotions since an increasing number of users express emotions via images and videos online. However, most existing methods based on convolutional neural networks aim to retrieve and classify affective images in a discrete label space while ignoring both the hierarchical and complex nature of emotions. On the one hand, different from concrete and isolated object concepts (*e.g.*, cat and dog), a hierarchical relationship exists among emotions. On the other hand, most widely used deep methods depend on the representation from fully connected layers, which lacks the essential texture information for recognizing emotions. In this work, we address the above problems via adaptive deep metric learning. Specifically, we design an adaptive sentiment similarity loss, which is able to embed affective images considering the emotion polarity and adaptively adjust the margin between different image pairs. To effectively distinguish affective images, we further propose the sentiment vector that captures the texture information extracted from multiple convolutional layers. Finally, we develop a unified multi-task deep framework to simultaneously optimize both retrieval and classification goals. Extensive and thorough evaluations on four benchmark datasets demonstrate that the proposed framework performs favorably against the state-of-the-art methods.**

***Index Terms*—Visual sentiment analysis, affective image retrieval, deep metric learning, convolutional neural network.**

I. Introduction

With the rapid development of digital technology and social networks, the scale of uploaded visual content has explosively grown. According to psychological theories [2], images can evoke various emotional stimuli for human viewers. It is significant and interesting to understand visual emotion due to its broad potential applications including emotion-based image retrieval (EBIR) [3], aesthetic quality categorization [4], and opinion mining [5], [6], [7], *etc*.

To date, many approaches have been proposed for classifying [8], [9] and retrieving [10], [11] affective images. There are several low-level features (*e.g.*, texture [8], color [12] and shape [13]) designed to represent the evoked emotion of visual contents. Recently, convolutional neural networks (CNNs) deliver end-to-end feature learning frameworks, which have been verified to outperform the hand-crafted features on emotion prediction tasks [14], [15].

Fig. 1. Examples from the Flickr and Instagram (FI) [14] datasets, which contain eight emotion categories as defined in Mikel's emotion wheel [16]. Obviously, the eight emotions are divided into two parts based on the polarity (*i.e.*, positive and negative). The emotions belonging to the same polarity are more related to each other.

Meanwhile, visual emotion analysis is challenging compared with conventional vision tasks (*e.g.*, object recognition and detection) due to the following two reasons: hierarchy and complexity of emotions. First, unlike concrete and specific object concepts (*e.g.*, dog, cat, and bird), emotions are not entirely independent of each other in terms of the semantic concept. As shown in Figure 1, there is an obvious hierarchy among the emotional label structure. There exist several fine-grained emotions in each of the polarities (*i.e.*, positive and negative). As the definition in Mikel's wheel, {amusement, contentment, awe, and excitement} belong to positive sentiment, while {fear, anger, disgust, and sadness} belong to negative sentiment. The emotions with the same polarity are more related in terms of the semantic concept. Nevertheless, existing studies always fail to consider the polarities of emotions for affective image analysis while they are trained in the isolated label space [17]. Second, most of the widely used CNN-based methods depend on the discriminative representation of deep features, especially the representation from fully connected (FC) layers [18]. There exists a latent assumption that the deep features, which perform well in distinguishing semantic information [19], can be utilized to accurately recognize abstract image emotions. However, since the emotion is an abstract concept, deep features lack interpretation for emotions [1] and may be insufficient to describe them. In fact, some literature has revealed that low-level information, such as texture, is an essential element to characterize visual emotion [8], [20]. We aim to capture essential texture information in feature embedding rather than directly using the representation from

X. Yao, D. She, H. Zhang, J. Yang and M.-M. Cheng are with the College of Computer Science, Nankai University, Tianjin 300350, China (e-mail: yxx_hbgd@163.com; sherry6656@163.com; zhhaiwei@nankai.edu.cn; yangjufeng@nankai.edu.cn; cmm@nankai.edu.cn).

L. Wang is with the National Laboratory of Pattern Recognition, CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: wangliang@nlpr.ia.ac.cn).

X. Yao and D. She contribute equally to this article.

A preliminary version of this work appeared at AAAI [1].

FC layers.

In this paper, we consider the hierarchy of emotion via deep metric learning [21], an approach that is good at constraining the distances between images in the separable space. However, compared with models optimized by softmax loss, the learned features acquired by the deep metric learning method may yield suboptimal results when exploited for classification. To address this problem, we integrate softmax loss and our proposed deep metric learning loss (*i.e*., adaptive sentiment similarity loss) into a unified multi-task framework. The learned features based on the framework can be simultaneously used for retrieval and classification tasks. In detail, the novel *adaptive sentiment similarity loss* is proposed to consider the hierarchical relationships among emotion categories, which is an extension of the triplet constraint. In the proposed loss function, the margin (*i.e*., the threshold value for the distance between dissimilar image pairs) can be adaptively adjusted according to the classification confidence from the softmax layer. In addition, inspired by the Gram matrix [22], we design a *sentiment vector* that captures sufficient texture information from multiple convolutional layers. The sentiment vector can model the correlations between different feature responses. We use the sentiment vector instead of the FC features to measure the similarity between different affective images, which is more effective and robust in characterizing emotions.

Our contributions are summarized as follows:

- We propose an adaptive sentiment similarity loss derived from triplet loss, taking the relation of emotion labels into consideration. With the assistance of softmax loss, we develop a multi-task framework to simultaneously optimize the retrieval and classification goals. Moreover, we can adjust the hyper-parameter *margin* in the adaptive sentiment similarity loss based on the classification confidence from the softmax layer.
- We design a sentiment vector capturing the texture information from multi-layers to measure the similarity between affective images in the embedding space. The experimental results on four widely used datasets verify that the proposed framework outperforms the state-of-art methods on both tasks.

Our new improvement compared with the preliminary conference version [1] lies in the following three aspects. (1) The framework is improved by adaptively adjusting the margin between image pairs in the similarity loss based on the classification confidence from the softmax layer. (2) More implementation details are provided, and sufficient visualization results are presented. Moreover, we systematically discuss the trade-off hyper-parameter ω in our combined loss function to investigate its sensitiveness. (3) We perform a more comprehensive survey of related work and provide a comparison with more contrastive methods in via experiments.

The rest of the paper is organized as follows. Section II summarizes the related work on affective image analysis and deep metric learning. Section III introduces the proposed method and the unified multi-task framework for affective image retrieval and classification. In Section IV, we present and visualize the experimental results on the popular benchmark datasets. Finally, Section V concludes this paper.

II. Related Work

In this section, we review the affective image analysis methods, including conventional hand-crafted features and CNN-based methods [23], [24], [25], and deep metric learning approaches [26], [27] that are most related to this work.

A. Affective Image Analysis

There are two typical models in the field of emotion studies, including dimensional emotion space (DES) and categorical emotion state (CES) models. DES models represent the sentiment in the two-dimensional valence-arousal (VA) coordinate space [28] or three-dimensional valence-arousal-dominance (VAD) space [29]. Zhao *et al*. [30] introduce a shared sparse regression model to predict the continuous probability distribution of image emotions in the VA space. By combining the different levels of deep features, Kim *et al*. [31] produce continuous emotion values in 2-D space. Meanwhile, CES models map emotions into one of the basic classes, such as *contentment* and *fear*, *etc*. Several typical models, including Ekman's six basic emotions [32] and Mikel's eight emotions [16], are used in numerous classification [15], [33] and discrete distribution prediction methods [34], [35]. Compared with DES, CES is easy to understand for users. Therefore, we analyze the emotions evoked by images using the CES model in this paper.

1) Hand-crafted Features: In early studies, many methods based on low-level features [8], mid-level representations [36], [37], [30] or high-level features [38] were developed. Low-level holistic features, such as Wiccest and Gabor features, are extracted to classify image emotions [39]. Later, in [40], Lu *et al*. provide an in-depth analysis for exploring the characteristic of shape features, empirically demonstrating that the proposed features can indeed capture emotions in the images. These low-level visual features are mainly proven to be effective on some small-scale datasets. Compared with low-level features, mid-level features such as attributes are more interpretable. They can bridge the gap between low-level features and high-level emotional information. To easily understand visual features, Wang *et al*. [41] propose interpretable aesthetic features to describe affective images, such as composition and color patterns. Additionally, various feature combinations inspired by the principle of art are designed in [9]. Further, Yuan *et al*. [42] introduce the Sentribute based on 102 mid-level attributes and facial expressions for affective image analysis, which is easy to understand and interpret for emotions. High-level features represent the semantic content in images. In [38], 1,200 adjective-noun pairs (ANPs) are developed to describe the sentiment semantic concepts. Each ANP can turn a natural noun into a concept that has strong emotion and is easy to detect. Based on multi-graph learning, Zhao *et al*. [11] explore the performance of affective image retrieval using different feature combinations, including high-level features from a semantic concept detector.

2) CNN-based Methods: Recently, with the superior capability of deep learning methods for feature learning [43], researchers have aimed to devise deep architectures for the affective image analysis. Many methods have been developed

for the classification task. DeepSentiBank [44] is constructed as a detector to describe the sentiment semantic concept based on 1,200 ANPs. This representation provides effective statistical cues for recognizing emotions evoked by images. Due to the limitation of training data, most methods aim to incorporate the model weights learned from a large-scale general dataset [45] and fine-tune the state-of-the-art CNNs on the target dataset for the task of image emotion prediction. Xu *et al*. [18] fine-tune the CNN pre-trained for object classification to recognize the emotions evoked by images. To use web data effectively to realize a more robust network, You *et al*. [17] develop a novel progressive strategy to learn the CNN framework. They select a potentially cleaner subset of web data to train the model iteratively. Liu *et al*. [46] use both hand-crafted features and semantic information generated by a CNN to train a classifier for affective images. The experimental results demonstrate that hand-crafted features can effectively supplement the deep features. Ragusa *et al*. [47] conduct comprehensive experiments to explore the effect of different CNNs and transfer learning on sentiment polarity recognition. To bridge the gap between different affective image datasets, the methods proposed in [48], [49] exploit transfer learning to leverage the learned information from the source domain to the target domain. To further obtain more discriminative features, several studies integrate local and global information into a final representation. Yang *et al*. [50] propose a weakly supervised coupled convolutional network with two branches. One branch is used to detect a sentiment soft map, while the other is used to couple the sentiment map with holistic deep features for classification. In [51], a multi-level region-based framework is designed to mine the local regions that evoke the emotional response for image classification. Fan *et al*. [52] evaluate the influence of focal attention on visual sentiment perception and design a DNN model that augmented with channels devoted to the feature of the focal region.

Most existing CNN-based architectures [53] aim to capture more useful information in the feature extraction process. However, in the feature embedding process, it is challenging to balance the complex distribution according to the hierarchy of emotion labels and the variations of intra- and inter- class. In this paper, we develop the adaptive sentiment similarity loss as the metric learning method considering the polarity concept in the embedding space. Then, we incorporate the similarity constraints and softmax loss into an end-to-end architecture. As shown in the experiments, this joint optimization strategy generates more discriminative and robust feature representations for both affective image retrieval and classification tasks.

B. Deep Metric Learning

Metric learning has been widely researched in pattern recognition and image analysis [54]. In the early period, conventional metric learning methods are mainly based on hand-crafted features. Recently, to directly and effectively measure the similarity between two images, deep metric learning has been exploited to map them into a separable space [55] via deep neural networks. Deep metric learning methods have been successfully applied to a diverse range of domains, *e.g*., face verification [56], [27], image retrieval [57], [58], and person re-identification [59], [60]. Many studies [21], [61], [27] employ CNNs with either pairwise (contrastive) [62] or triplet constraints [63] to learn feature embeddings capturing the semantic similarity among images. To reduce the heavy computational burden, N-pair loss [64] exploits an effective batch construction strategy in the training process. Lifted structure embedding [65] takes full advantage of the training batches by extending the pairwise distances in a mini-batch to the dense matrix. Instead of using similarity constraints alone, some strategies incorporate softmax loss into discriminative feature generation [57], [66]. Center loss [66] is proposed to minimize the intra-class distances of the deep features. Combining center loss and softmax loss can enhance the discriminative power of features for face recognition. In [67], He *et al*. design a triplet-center loss to minimize the intra-class variations and simultaneously maximize the inter-class distances, while softmax loss is used as an auxiliary loss. Different from the methods based on discrete labels, Kim *et al*. [68] explore a novel metric learning method based on continuous and structured labels. The method can be used for image retrieval tasks where the label of data is continuous, such as human poses.

A large fraction of trivial pairs selected randomly will not contribute to the loss and gradient. Therefore, to achieve better performance and faster convergence during the training process, quite a few studies focus on the sample mining method [69], [70]. By combining the triplet model and the global structure of the embedding space, a smart mining procedure is introduced in [71]. In addition, Duan *et al*. [26] design a novel deep adversarial learning method to generate hard negatives from easy negatives to train a more robust model. Zheng *et al*. [72] exploit linear interpolation to adaptively manipulate the hard level of the synthetic training data to control the training process.

In this paper, based on triplet loss, the adaptive sentiment similarity loss is introduced by considering the hierarchical relation in emotions. Other than adopting an essential sampling method proposed in [70], we also adaptively adjust the margin between different samples.

III. Methodology

We design a unified multi-task framework that is simultaneously optimized by the adaptive sentiment similarity loss and softmax loss for affective image analysis. Figure 2 shows the pipeline of the proposed architecture. The designed sentiment vectors are extracted from multi-layers for adaptive feature embedding.

A. Problem Formulation

In this work, our goal is to simultaneously retrieve and classify affective images through a unified framework. The framework is trained with the joint supervision of the metric learning loss (*i.e*., adaptive sentiment similarity loss) and classification loss (*i.e*., softmax loss). Given N training images $\{(x_i, y^{x_i})\}_{i=1}^{N}$ of C categories, x_i denotes the i^{th} training image, and $y^{x_i} \in \{1, 2, \cdots, C\}$ represents the corresponding

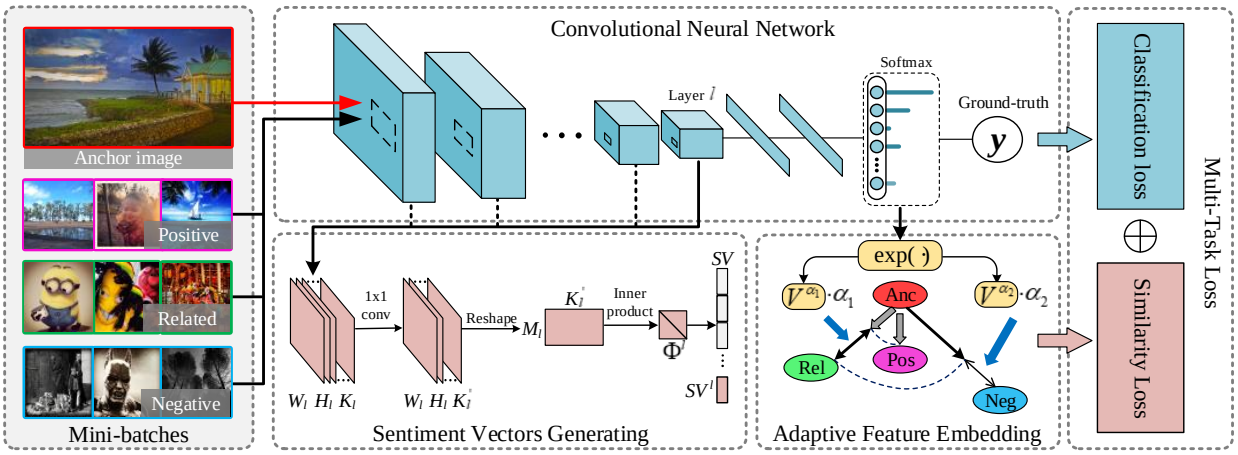


Fig. 2. Pipeline of the proposed approach. Given the mini-batches containing images with different emotions, we first generate the sentiment vector from the Gram matrix in the convolutional layer to measure the distances of different affective images and then adaptively learn feature embedding based on the confidence scores in the softmax layer. Our framework is optimized with the joint supervision of the classification loss (*i.e.*, softmax) and similarity loss (*i.e.*, adaptive sentiment similarity loss). In layer l, the dimension of each 3-D feature map is $W_l \times H_l \times K_l$, and $M_l = W_l \times H_l$. The Gram matrix is represented by $\Phi^l \in \mathbb{R}^{K_l' \times K_l'}$, and SV denotes the sentiment vector. Anc, Rel, Pos, and Neg represent the anchor example, related example, positive example, and negative example, respectively. The weight of margin α_1 is denoted by V^{α_1}, while the weight of margin α_2 is denoted by V^{α_2}.

emotion labels. Based on the CNN $\mathcal{F}(\cdot, \theta)$ (θ represents the parameters of the network), these images are embedded into feature vectors by the metric learning loss, which balances the complicated distribution of intra- and inter- class variations. Then, for image x_i, we maximize the confidence score $\boldsymbol{c}^{x_i}$ for the ground-truth y^{x_i} by optimizing the classification loss, which enforces that a distance is maintained between the deep features of different classes. By leveraging the advantages of the metric learning loss and classification loss, we can obtain more discriminative features for both retrieval and classification tasks in the unified end-to-end framework.

B. Sentiment Metric Learning

We measure the similarity between affective images by relying on the Euclidean distance D of their feature embeddings that contain texture information with the unit norm:

$$D(x_i, x_j) \mapsto \|SV_i - SV_j\|_2 . \tag{1}$$

x_i and x_j refer to the anchor images from the training set Γ, while SV_i and SV_j denote the corresponding sentiment vectors containing texture information extracted from multi-level convolutional layers. Although the distances between different affective images are subjective, the general relationship is clear and easy to define. According to the polarity concept involved in emotion labels, the following goal should be well satisfied: images of the same polarity are closer to each other compared with images of the opposite polarity in the embedding space. Therefore, we extend the triplet constraint to the sentiment constraint by taking into account the emotion polarity.

1) Review on Triplet Constraints: Existing methods [70] based on triplet constraints always generate mini-batches of triplets, *i.e.*, an anchor a_i, a positive example p_i of the same class, and a negative example n_i of a different class. The goal is to pull examples belonging to the same class into nearby points in a manifold space and push examples belonging to different classes apart from each other, which can be expressed as:

$$D(a_i, p_i) + \alpha < D(a_i, n_i), \forall (a_i, p_i, n_i) \subset \Gamma, \tag{2}$$

where $\alpha > 0$ denotes the margin that controls the minimum distance between negative and positive examples. The distances between examples from the same class (*i.e.*, a_i and p_i) should be smaller than those from different classes (*i.e.*, a_i and n_i) by at least a margin α.

2) Sentiment Similarity Loss: In general, triplet loss is extensively used as a metric learning method in various tasks. However, there exists a hierarchical relation in the emotion structure, which cannot be directly taken into account in triplet loss. Therefore, it is essential to differentiate examples of different classes with an anchor based on their polarity when learning feature embedding. For instance, Mikel's eight emotions [16] have four positive (*i.e.*, amusement, contentment, awe, and excitement) and four negative (*i.e.*, fear, sadness, disgust, and anger) emotions. For two emotion categories with the same polarity, one is the related example r_i of the other. Given an anchor image a_i of a specific emotion, we ensure that the positive example p_i of exactly the same emotion is closer to a_i than the related example r_i. Meanwhile, the negative example n_i of the opposite polarity remains the farthest distance away. Therefore, we propose the sentiment constraint, which can be defined as the following formulation:

$$\begin{cases} D(a_i, p_i) + \alpha_1 < D(a_i, r_i) \\ D(a_i, r_i) + \alpha_2 < D(a_i, n_i) \end{cases}, \forall (a_i, p_i, r_i, n_i) \subset \Gamma, \tag{3}$$

where $\alpha_1, \alpha_2 > 0$ represent the margins between different emotion labels. We show the difference between the triplet constraint and sentiment constraint in Figure 3. Based on the sentiment constraint, we propose to minimize the sentiment

similarity loss function:

$$L_{stm} = \sum_{i=1}^{N} [D(a_i, p_i) - D(a_i, r_i) + \alpha_1]_+ + \sum_{i=1}^{N} [D(a_i, r_i) - D(a_i, n_i) + \alpha_2]_+ , \quad (4)$$

where $[\cdot]_+ = \max(0, \cdot)$. N denotes the number of affective images in the training set.

3) Adaptive Sentiment Similarity Loss: Given an image x_i, each confidence score $\boldsymbol{c}_j^{x_i} \in [0, 1]$ can be regarded as the degree of the tendency toward class j. In this paper, we expect that the confidence score on the given ground-truth will be approximately 1. Therefore, we apply a stronger penalization on samples that have a high confidence score for classes other than the ground-truth.

Given a sampled quadruplet consisting of the anchor, positive, related, and negative examples, we aim to adaptively adjust the margins (*i.e.*, α_1 and α_2) in the sentiment similarity loss guided by the confidence score from the classifier. For α_1, which controls the margin between the anchor-positive pair and anchor-related pairs, we weight it according to the anchor-related pair (a_i, r_i). Specifically, for an anchor a_i from the y^{a_i} class and related example r_i from the y^{r_i} class, a higher confidence of a_i w.r.t. the y^{r_i} class or r_i w.r.t. the y^{a_i} class denotes that the pair is harder to separate. Consequently, we assign a higher weight on α_1 resulting in a stronger penalization in the process. We weight α_2, which controls the margin between the anchor-related pair and anchor-negative pair, according to the similarity of the anchor and negative examples guided by the confidence score. Specifically, $\boldsymbol{c}_{y^{r_i}}^{a_i}$ denotes the confidence of anchor a_i w.r.t the class of the related example r_i. Analogously, we can define $\boldsymbol{c}_{y^{a_i}}^{r_i}$, $\boldsymbol{c}_{y^{n_i}}^{r_i}$, and $\boldsymbol{c}_{y^{r_i}}^{n_i}$. The weights are formulated as follows:

$$v^{\alpha_1} = \exp(\boldsymbol{c}_{y^{r_i}}^{a_i}) \cdot \exp(\boldsymbol{c}_{y^{a_i}}^{r_i}), \quad (5)$$

$$v^{\alpha_2} = \exp(\boldsymbol{c}_{y^{n_i}}^{r_i}) \cdot \exp(\boldsymbol{c}_{y^{r_i}}^{n_i}), \quad (6)$$

where v^{α_1} and v^{α_2} represent the weights of α_1 and α_2, respectively, for specific quadruplet examples. The values of v^{α_1} and v^{α_2} for each quadruplet are adaptively adjusted according to their classification confidence.

Therefore, the adaptive sentiment similarity loss can be formulated as follows:

$$L_{astm} = \sum_{i=1}^{N} [D(a_i, p_i) - D(a_i, r_i) + v^{\alpha_1}\alpha_1]_+ + \sum_{i=1}^{N} [D(a_i, r_i) - D(a_i, n_i) + v^{\alpha_2}\alpha_2]_+ . \quad (7)$$

Here, L_{astm} denotes the adaptive sentiment similarity loss, which is the extension of L_{stm}.

C. Sentiment Vector for Feature Embedding

While many recent studies aim to extract more discriminative high-level semantic features due to the powerful representation capacity of deep learning methods, low-level features such as texture are not well considered in the field of visual emotion. It is has been verified that texture is one of the crucial low-level features for describing image emotion [8], [73]. To extract informative texture representations, we design the sentiment vector derived from Gram matrices in multi-layers. Each element in a Gram matrix represents the inner product of different filter responses in a convolutional layer [22]. Therefore, we can obtain both high-level FC features and low-level representations, especially texture, in the framework.

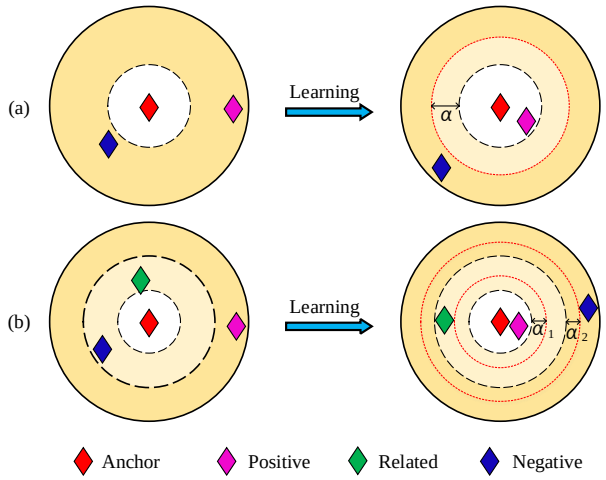


Fig. 3. Compared with the triplet constraint (a) considering anchor, positive, and negative examples, our sentiment constraint (b) consisting of the anchor, positive, related, and negative examples takes the hierarchy of emotion labels into account. Given an anchor image, α represents the margin between the anchor-positive pair and anchor-related pair, while α_1 represents the margin between the anchor-related pair and anchor-negative pair. α_2 represents the margin between the anchor-positive pair and the anchor-negative pair.

Specifically, we first compute the activation of each convolutional layer $l \in \{1, 2, \cdots, L\}$ for every image fed into the network. Now that each layer in the network can be regarded as a non-linear filter bank, its response to an image organizes a series of feature maps. Suppose that there are K_l filters in layer l, so it will generate K_l feature maps. The size of each 2-D feature map is denoted as $W_l \times H_l$, where W_l and H_l represent the width and height, respectively. In a convolutional layer, many filters result in thousands of elements in the Gram matrix, which leads to massive computational consumption. To reduce the computational burden of the proposed framework, a 1x1 convolution layer is added after feature maps to shrink the size of the Gram matrix. Through the process, the number of 2-D feature maps from layer l is reduced from K_l to K'_l. In layer l, the response of the j^{th} feature maps at position m is represented as F_{mj}^l, which forms the matrix $F^l \in \mathbb{R}^{M_l \times K'_l}$. K'_l is the number of feature maps in layer l after dimension reduction and $M_l = W_l \times H_l$. These feature maps can be aggregated into a 2-D Gram matrix $\Phi^l \in \mathbb{R}^{K'_l \times K'_l}$, which discards the spatial information that is not needed for stationary texture. Each element Φ_{ij}^l of the Gram matrix in layer l is obtained by computing the inner product between the i^{th} and

Algorithm 1 Training the multi-task framework for affective image retrieval and classification.

Input:

The training set $\Gamma = \{x_1, x_2, ..., x_N\}$

Output:

The learnable parameters θ of the multi-task neural network $\mathcal{F}(\cdot, \theta)$

Initialize the framework $\mathcal{F}(\cdot, \theta)$ with a pre-trained CNN.
repeat
  Feed training images through the CNN model from the first layer to the last convolutional layer.
  Compute sentiment vector (SV^l) generated from the Gram matrix in convolutional layer l.
  Concatenate SV^ls from each convolutional layer and normalize them using the l_2 norm.
  Sample quadruplet images (a_i, p_i, r_i, n_i), and compute their confidences for each category in the classifier.
  Adaptively weight the margins α_1 and α_2 based on confidence scores.
  Compute the total loss, summing the classification loss and adaptive sentiment similarity loss.
  Backpropagate the gradients, and update the learnable parameters.
until Converge

j^{th} vectorized feature maps:

$$\Phi_{ij}^l = \sum_{m=1}^{M_l} F_{mi}^l F_{mj}^l, \tag{8}$$

so each element captures the correlation between two feature maps.

It is obvious that the Gram matrix is symmetrical, and thus the number of independent elements is $K_l'(K_l'+1)/2$. The sentiment vector SV from layer l is designed as follows:

$$SV^l = [\Phi_{1,1}^l, \Phi_{2,1}^l, \Phi_{2,2}^l, \cdots, \Phi_{K_l',1}^l, \cdots, \Phi_{K_l',K_l'}^l]. \tag{9}$$

We concatenate the sentiment vector from each layer as $SV = [SV^1, SV^2, \cdots, SV^L]$. Then the combined sentiment vector SV is normalized to a unit l_2 norm to form the feature embedding used in sentiment metric learning.

D. Multi-Task Framework

In the traditional training process for a recognition task, softmax loss is extensively utilized to maximize the probability of the ground-truth, and its effectiveness has been validated in many studies. The activation value of unit j for image x_i in the last FC layer is denoted as $h_j^{(i)}$, where $j \in \{1, 2, \cdots, C\}$. The probability of $y^{x_i} = j$ is denoted by $\boldsymbol{c}_j^{x_i}$, representing the confidence toward the corresponding class:

$$\boldsymbol{c}_j^{x_i} = \frac{\exp(h_j^{(i)})}{\sum_{k=1}^{C} \exp(h_k^{(i)})}. \tag{10}$$

Then, the network is optimized by minimizing softmax loss:

$$L_{cls} = -\frac{1}{N}\left[\sum_{i=1}^{N}\sum_{j=1}^{C} \mathbf{1}(y^{x_i} = j) \ln \boldsymbol{c}_j^{x_i}\right], \tag{11}$$

where $\mathbf{1}(\delta) = 1$ if the condition δ is true; otherwise, it is 0.

Softmax loss can be regarded as the sum of the negative log-likelihood on the whole training dataset $\{x_i\}_{i=1}^N$. It penalizes the classification error for each category equally and thus overlooks the hierarchy of emotion labels. Therefore, given the quadruplets and the labels of images as input, we explicitly train the unified framework by optimizing the classification and adaptive sentiment similarity losses. While softmax loss cannot shorten the intra-class distance, the similarity loss can effectively supplement it. Meanwhile, softmax loss can guide the similarity loss to differentiate different classes. The total loss function is integrated with two losses via a weighted combination:

$$L_{total} = (1-\omega)L_{cls} + \omega L_{astm}, \tag{12}$$

where ω is the weight to control the trade-off between the two loss functions. Algorithm 1 summarizes the overall learning procedure of the proposed multi-task framework.

IV. Experimental Results

In this section, we conduct extensive and thorough experiments to validate the effectiveness of our multi-task framework for affective image analysis. Specifically, based on the sentiment vector from Gram matrices, our learned feature embeddings for affective image retrieval outperform the state of the art on four datasets. Meanwhile, our framework also achieves promising classification performance compared with several baseline methods.

A. Datasets

We conduct the experiments on four datasets, including Flickr and Instagram (FI) [14], Subset A of IAPS (IAPSa), the Artistic dataset (ArtPhoto) and Abstract Paintings (Abstract) [8]. FI is obtained from Flicker and Instagram by querying with Mikel's eight emotions as keywords resulting in 90,000 noisy images. Ultimately, a total of 23,308 images receive at least three agrees according to the labels annotated by 225 Amazon Mechanical Turk workers. [1] The International Affective Picture System (IAPS) [74] is an emotion evoking image set in the psychology domain that is widely exploited in affective image analysis research, from which IAPSa is constructed, which includes 395 images annotated with the same eight emotion categories. ArtPhoto contains 806 artistic photographs from a photo sharing site, whose emotion labels rely on the artist that uploads the photo. Abstract consists of 228 peer rated abstract paintings that contain plentiful color and texture characteristics.

B. Implementation Details

The developed framework is based on GoogLeNet-Inception-v3 [75] rather than Inception-v1 [76] used in our earlier work [1]. First, we transfer the weights trained on the large-scale dataset [45] to the framework, and then fine-tune on

[1]We have 22,713 manually labeled images as some images no longer exist on the Internet.

TABLE I
RETRIEVAL AND CLASSIFICATION PERFORMANCE ON THE FI DATASET. WE COMPARE VARIOUS BASELINES FOR LEARNING THE EMOTION REPRESENTATION, INCLUDING THE CONVENTIONAL ALGORITHMS AND CNN-BASED METHODS. HERE, 'S + T' INDICATES UTILIZING BOTH SOFTMAX AND TRIPLET LOSSES TO JOINTLY TRAIN THE FRAMEWORK. THE FEATURES FROM THE FULLY-CONNECTED (FC) LAYER ARE EXTRACTED IN COMPARED METHODS, AND 'DIM.' REPRESENTS THE DIMENSION OF FEATURES. OUR METHOD IS BASED ON THE INCEPTION-V3 ARCHITECTURE.

Algorithms	Dim.	Acc.(%)	Retrieval Performance						
			mAP_8 ↑	mAP_2 ↑	FT ↑	ST↑	NN ↑	DCG↑	ANMRR↓
SIFT	1,000	37.56	0.1705	0.5913	0.1830	0.3513	0.2462	0.4507	0.6553
HOG	1,000	44.67	0.2115	0.6002	0.1926	0.3620	0.3225	0.4639	0.6424
Gabor	1,000	36.33	0.1724	0.5942	0.1768	0.3395	0.2641	0.4434	0.6770
ORB	1,000	38.12	0.1824	0.6154	0.1798	0.3533	0.2332	0.4523	0.6889
SentiBank	1,200	49.09	0.2337	0.6168	0.2422	0.4232	0.3990	0.5223	0.5934
DeepSentiBank	2,089	56.15	0.2559	0.6247	0.2658	0.4468	0.4583	0.5509	0.5655
MVSO	4,342	60.36	0.2798	0.6366	0.2877	0.4761	0.5158	0.5158	0.5731
AlexNet (Softmax)	4,096	58.13	0.2709	0.6328	0.2795	0.4693	0.5038	0.5633	0.5463
VGGNet (Softmax)	4,096	64.55	0.3013	0.6552	0.3007	0.4887	0.5511	0.5860	0.5161
Inception-v1 (Softmax)	1,024	65.18	0.3583	0.6773	0.3571	0.5619	0.5816	0.6403	0.4517
Inception-v3 (Softmax)	2,048	65.52	0.5756	0.7336	0.5429	0.6915	0.6399	0.7431	0.3154
Triplet (Inception-v1)	1,024	63.46	0.3951	0.6981	0.3932	0.6081	0.5578	0.6762	0.4082
Triplet (Inception-v3)	2,048	64.38	0.6489	0.7652	0.5970	0.7236	0.6168	0.7667	0.2782
S + T (Inception-v1)	1,024	65.35	0.4426	0.7592	0.4435	0.6513	0.5866	0.7119	0.3603
S + T (Inception-v3)	2,048	66.83	0.6794	0.7837	0.6352	0.7715	0.6391	0.8007	0.2350
CroW (Inception-v3)	2,048	62.06	0.6121	0.7574	0.5731	0.7224	0.6501	0.7671	0.2861
R-MAC (Inception-v3)	2,048	65.34	0.6331	0.7633	0.5935	0.7462	0.6684	0.7873	0.2606
Center loss (Inception-v3)	2,048	61.28	0.6216	0.7238	0.5912	0.6944	0.6036	0.7496	0.3034
N-pair loss (Inception-v3)	2,048	66.12	0.6574	0.7727	0.6227	0.7332	0.6210	0.7823	0.2535
Binomial deviance (Inception-v3)	2,048	66.57	0.6748	0.7740	0.6573	0.7166	0.6362	0.7665	0.2789
Lifted structure (Inception-v3)	2,048	66.89	0.6891	0.7444	0.6557	0.7230	0.6452	0.7693	0.2738
FastAP (Inception-v3)	2,048	61.15	0.6731	0.7759	0.6293	0.7480	0.5989	0.7805	0.2589
SoftTriple (Inception-v3)	2,048	64.52	0.7008	0.7891	0.6677	0.7703	0.6518	0.7964	0.2347
Ours (Inception-v3)	680	**68.37**	**0.7319**	**0.9192**	**0.6966**	**0.7892**	**0.6727**	**0.8197**	**0.2100**

the FI dataset. The dataset is split randomly into 80% training, 5% validation and 15% test sets. During the training process, the original images are resized to 356×356, followed by a center 299×299 cropping. We initialize the learning rate as 10^{-3}, and reduce it one-tenth every 30 epochs. We fine-tune all layers by stochastic gradient descent (SGD) throughout the whole net using batches of 32. A total of 100 epochs are run to update the parameters, which is sufficient for our framework to converge. The margin α in triplet loss is set to 0.2, and the base margins α_1 and α_2 in the proposed similarity loss are set to 0.2 and 0.1, respectively. We set the weight ω as 0.2, and detailed discussions regarding its sensitivity are given in the following subsection. The feature dimension for triplet loss is 2048. The dimension of the sentiment vector from the last convolutional layer in each stage is 136, using 16 filters with a kernel size of 1×1. After concatenating the vectors from five stages, we obtain a total of 680-dimensional features as embeddings. We conduct all our experiments on two NVIDIA GTX 1080 GPUs. In addition, we also evaluate our framework on three small datasets that have limited training images with the assistance of transfer learning. Specifically, we fine-tune the framework trained on the FI dataset using the small-scale datasets, which are randomly split into 80% training and 20% test sets. We conduct 5-fold validation and report the average performance.

C. Baseline

In this subsection, we evaluate the proposed framework against the state-of-the-art methods for affective image analysis including retrieval and classification. We compare our framework with different algorithms, including methods using hand-crafted features and deep features.

1) Hand-crafted Features: We extract three low-level features including local descriptors such as SIFT [77], HOG [78], Gabor [79] and ORB [80]. We also use a concept detector library based on the constructed ontology, named SentiBank [38], to utilize the 1,200-dimensional ANP features as the high-level representation.

2) Deep Models: The 2,089-dimensional features of DeepSentiBank [38] and 4,342-dimensional features of MVSO [81] are exploited as deep features. In addition, we pay attention to the performance of deep features of CNN models trained with different constraints, and a variety of architectures are also evaluated in our experiments, *i.e.*, AlexNet [82], VGGNet [83], GoogLeNet-Inception-v1 [76], and GoogLeNet-Inception-v3 [75]. All of the depicted models are initialized with weights from architectures pre-trained on the ImageNet and fine-tuned on the affective datasets, where softmax loss is employed for optimization. We report the results of using features extracted from the last FC layer. Meanwhile, we also present the performance when using feature postprocessing methods, such as CroW [84] and R-MAC [85], which are proposed for image retrieval. The feature dimensions of different architectures are shown. Additionally, we present the performance of the CNN that directly employs the similarity loss (*e.g.*, triplet loss [70], N-pair loss [64], binomial deviance [86], lifted structure loss [65], SoftTriple [87], and FastAP [88]) for learning the affective image representation, and the model

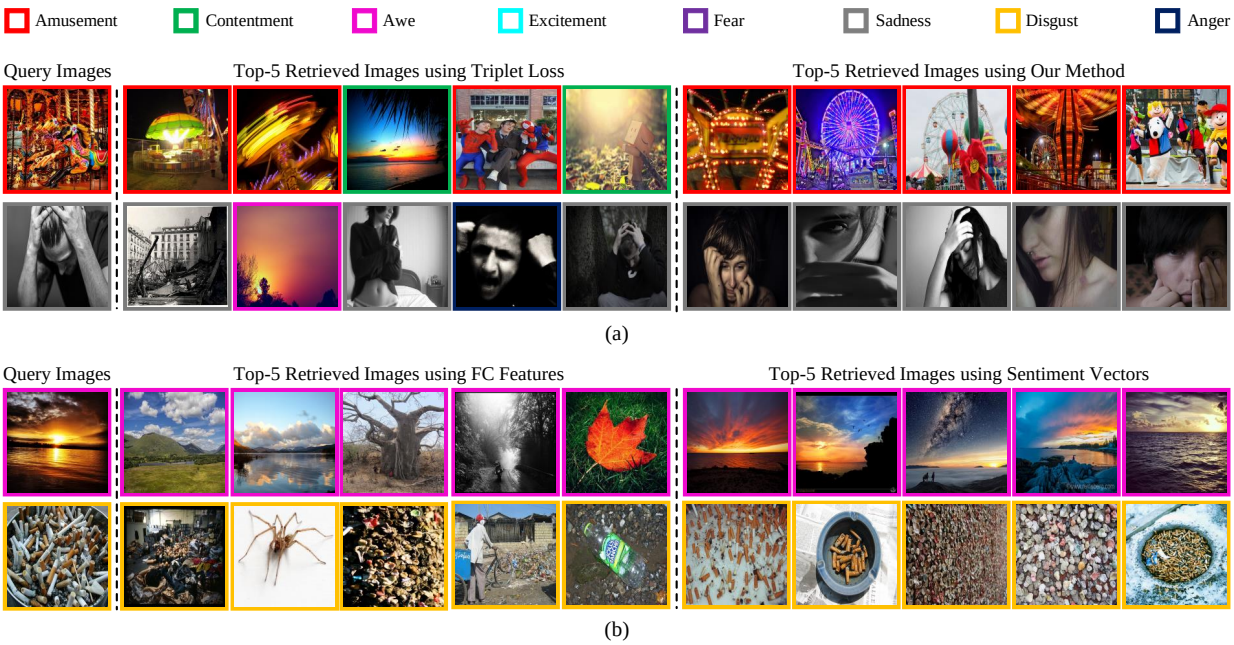


Fig. 4. Top-5 retrieval results of sampled query images (first column) using the different methods on the FI dataset. The images in different color boxes belong to different emotion categories. (a) Retrieval results of the network trained by triplet loss and our proposed framework. (b) Retrieval results obtained using the fully-connected (FC) features learned by our proposed similarity loss and using the sentiment vectors learned by our proposed similarity loss. Based on sentiment vectors, the images that have texture information similar to the query are easily retrieved.

jointly optimized by softmax loss and triplet loss [70].

D. Evaluation

Given a query image, we expect to retrieve the images that have a similar emotion, which is measured by the Euclidean distance of the feature representations in the embedding space. For the FI dataset, each image in the test set is used as the query to search the samples with a similar emotion among the training images. Following [11], we use each image to retrieve the positive samples from the remaining ones when evaluating the framework on the three small datasets. We employ the following universal measurements for the retrieval task in our experiments. Nearest neighbor rate (NN) calculates the precision of the sample that ranks first in the returned list. First tier (FT) and second tier (ST) are defined as the recall of the top-n and top-$2n$ returned results, respectively, where n denotes the number of relevant images in the retrieval database. The mean precision of the retrieval results is defined as the mean average precision (mAP). In detail, we consider the mean precision of eight specific emotion categories (*i.e.*, mAP_8) and the mean precision of the two polarities (*i.e.*, mAP_2) simultaneously. Since the frontal results are given more attention, the discounted cumulative gain (DCG) is designed to measure the results with a higher weight appended to positions that are more frontal. $F1$ is a measure of accuracy and is defined based on the values of the precision and recall. Average normalized modified retrieval rank (ANMRR) considers the ranking order of relevant samples in the returned results. All the measurements range from 0 to 1. A lower value of ANMRR indicates better performance, while a higher value of the other measurements represents better performance.

E. Results on Large-Scale Datasets

First, we evaluate the proposed method against different algorithms on the most popular and largest manually annotated dataset (*i.e.*, FI) [14].

1) Affective Image Retrieval: We conduct experiments for evaluating the retrieval performance using different methods on the test set of the FI dataset. As shown in Table I, the low-level generic local features, including SIFT, HOG, and Gabor, obtain poor performance due to the lack of capacity to represent emotions. The representations extracted from SentiBank, DeepSentiBank and MVSO achieve similar results. They only outperform the performance of low-level features by a small margin. The CNN models trained for affective image classification exhibit significantly improved retrieval performance compared with those trained for object classification, since the models have learned affective-level discriminative representations. It is observed that the more advanced Inception-v3 architecture pre-trained for affective image classification achieves much better performance than Inception-v1, which was utilized in our earlier work [1]. After feature weighting and pooling, Crow and R-MAC exhibit improved performance on all the metrics compared with the FC feature learned only by softmax loss. Moreover, a metric learning method, such as triplet loss, is shown to perform better than softmax loss on affective image retrieval. This is mainly because metric learning focuses more on the distance between feature points in the embedding space.

With the joint supervision of softmax loss and triplet loss, the model can learn features guided by the goals of inter-class dispersion and intra-class compactness. The effectiveness in the retrieval task is further promoted. By introducing the

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.2020.3001527, IEEE Transactions on Multimedia

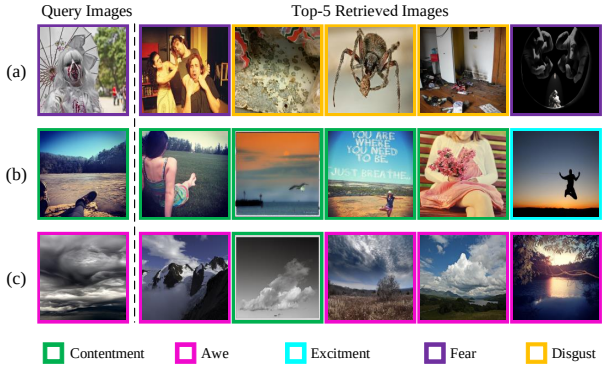


Fig. 5. Failure cases arising in the top-5 retrieved images. The images in different color boxes belong to different emotion categories.

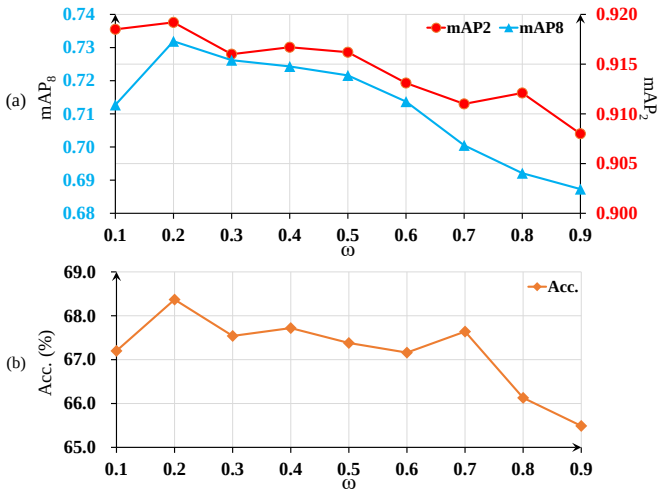


Fig. 6. Discussion of hyper-parameter ω in the objective loss function. The experiment is conducted on FI dataset. (a) The results on 'mAP_8' and 'mAP_2' metrics for retrieval task. (b) The accuracy (Acc.) of classifying affective images.

sentiment vector capturing texture information as well as the adaptive sentiment similarity loss, our proposed method improves $\mathbf{mAP_8}$ and $\mathbf{mAP_2}$ by a large margin. Additionally, we also report the performances of the networks optimized by the recently proposed metric loss functions (*i.e.*, N-pair loss, center loss, binomial deviance, lifted structure loss, SoftTriple, and FastAP), which also fail to outperform the combined loss consisting of softmax and our adaptive sentiment similarity losses.

We also show our top-5 retrieval results from the FI dataset in Figure 4, which demonstrate the obvious effectiveness of our designed framework for affective image retrieval. As shown in Figure 4(a), our method has obvious superiority compared with directly using triplet loss. For the first two query images, our method can clearly measure the similarity between affective images based on high-level semantics. However, the network only based on triplet loss obtains incorrect results in the retrieved top-5 images, although their hues are close to those in the query images. Additionally, there are images of opposite polarity to the query in the top-5 retrieval results. In contrast, the similarity loss considering the hierarchical structure in emotion , as well as softmax loss, improves the discriminant ability of features. Figure 4(b) presents the retrieval results obtained using FC features and sentiment vectors based on our proposed combined loss function. Although the top-5 retrieved images are all correct, we can retrieve an image whose texture information is closer to the query based on the proposed sentiment vectors compared with FC features.

In Figure 5(a), there are three failure cases in the top-5 retrieved samples for the query image. After viewing the query, both fear and disgust emotions are evoked in people, which leads to retrieved images from the 'disgust' category. In fact, a single label may be insufficient to describe the emotion in an image, which is called the ambiguity of emotion. In the future, a significant direction will therefore be studying how to retrieve images based on the similarity of an emotion distribution rather than a single label. For the query image in Figure 5(b) from the 'contentment' category, the retrieved image ranking fifth belongs to the 'excitement' category. We can observe that the background of the retrieved image evokes the contentment emotion, while the jumping person presented in a small region of the image evokes the excitement emotion. However, the jumping person could draw more viewers' attention compared with the background. We surmise that the attention mechanism can be introduced into our method to tackle this problem in the future. The label of the retrieved image ranking second in Figure 5(c) is contentment, but it has high similarity with the query. It might be mislabeled according to our observation, so high-quality affective datasets are also desired in the field of visual emotion analysis.

2) Affective Image Classification: We also present the classification results of different methods on the FI dataset in Table I. The CNN-based deep features outperform the hand-crafted local descriptors since they can learn a more discriminative representation to classify image emotions. Meanwhile, the models with a deeper structure can achieve better performance. The model optimized with only triplet loss obtains worse results than those fine-tuned by softmax loss, as the deep metric loss is more adequate for retrieval than it is for classification. However, softmax loss and the triplet loss can promote each other, resulting in the more discriminative framework, in which the accuracy improves by approximately 1%. Our algorithm further outperforms the fine-tuned CNN models by an approximately 3% improvement.

3) Influence of Hyper-parameters: We systematically analyze the influence of hyper-parameter ω in Equation 12, where ω controls the relative importance between the similarity loss and softmax loss. In the optimization function, the larger the value of ω is, the more important the similarity constraint term is. We illustrate how ω influences the performance of the total loss using three metrics on the FI dataset (*i.e.*, 'Accuracy' for classification and 'mAP_8' and 'mAP_2' for the retrieval task). Since the two losses are mutually related, we consider the results for ω ranging from 0.1 to 0.9. This ensures that the two constraints simultaneously exist in our framework. As shown in Figure 6, we present the performance in terms of the three metrics when ω is set to different values. From the results, we can make the following observations: (1) Compared with

TABLE II

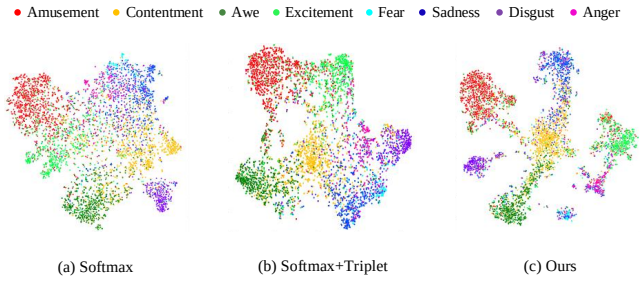
ABLATION STUDY ON THE FI DATASET. HERE, '$*$', '$\bullet$' AND '$\diamond$' DENOTE USING DIFFERENT SAMPLING METHODS (*i.e.*, RANDOM SAMPLING, HARD SAMPLING, AND SEMI-HARD SAMPLING). 'SENTI' DENOTES SENTIMENT SIMILARITY LOSS, WHILE 'ASENTI' DENOTES ADAPTIVE SENTIMENT SIMILARITY LOSS. DIFFERENT EMBEDDINGS ARE EMPLOYED TO REPRESENT SENTIMENTS, *i.e.*, FULLY CONNECTED (FC) LAYERS AND SENTIMENT VECTORS (SVS). THE RESULTS IN BLUE FONT IN THE PENULTIMATE ROW ARE OBTAINED USING THE METHOD IN OUR CONFERENCE VERSION, WHILE RESULTS IN RED FONT IN THE LAST ROW ARE OBTAINED USING OUR NEW METHOD PROPOSED IN THIS PAPER.

Loss Function				Feature	Acc.(%)	mAP$_8$ ↑	mAP$_2$ ↑	FT ↑	ST ↑	NN ↑	DCG ↑	ANMRR ↓
Softmax	Triplet	Senti	Asenti									
√				FC	65.52	0.5756	0.7336	0.5429	0.6915	0.6399	0.7431	0.3154
	√*			FC	64.38	0.6489	0.7652	0.5970	0.7236	0.6168	0.7667	0.2782
√	√*			FC	66.21	0.6687	0.7813	0.6227	0.7654	0.6564	0.7994	0.2386
√	√$^\bullet$			FC	66.14	0.6726	0.7840	0.6316	0.7666	0.6554	0.7979	0.2377
√	√$^\diamond$			FC	66.83	0.6794	0.7837	0.6352	0.7715	0.6391	0.8007	0.2350
√		√$^\diamond$		FC	67.40	0.6908	0.8956	0.6453	0.7696	0.6406	0.8071	0.2247
√		√$^\diamond$		SV	67.82	0.7138	0.9083	0.6739	0.7759	0.6602	0.8127	0.2221
√			√*	SV	67.89	0.7259	0.9135	0.6808	0.7841	0.6713	0.8088	0.2118
√			√$^\bullet$	SV	67.65	0.7245	0.9079	0.6713	0.7821	0.6666	0.8070	0.2133
√			√$^\diamond$	SV	68.37	0.7319	0.9192	0.6966	0.7892	0.6727	0.8197	0.2100

TABLE III

MAP$_8$ RESULTS OBTAINED USING A FIXED MARGIN AND AN ADAPTIVE MARGIN. THE EXPERIMENTS ARE INDEPENDENTLY REPEATED FIVE TIMES.

Method	#1	#2	#3	#4	#5
Fixed margin	0.7062	0.7140	0.7236	0.7190	0.7138
Adaptive margin	0.7253	0.7284	0.7323	0.7265	0.7319



(a) Softmax (b) Softmax+Triplet (c) Ours

Fig. 7. Visualization of feature embeddings optimized by different constraints. Each point represents an image in the test set of FI. Different emotion labels are distinguished by different colors. (a) and (b) show the feature space when using the FC feature learned by softmax loss and the combined loss respectively, while (c) presents the feature space when using the sentiment vector learned from our framework. It is observed that our framework can separate the images from different emotion categories more effectively.

'mAP$_2$', 'mAP$_8$' is more sensitive to changes in the value of ω. (2) The performance on the two tasks is relatively better when softmax loss is assigned a higher weight, which demonstrates that softmax loss is the guidance for the training process, and then, similarity constraints can optimize the framework better. (3) The best results in terms of the three metrics are simultaneously obtained when $\omega = 0.2$ (*i.e.*, , the weight of softmax loss is 0.8).

In general, the proposed method is robust for both the affective image retrieval and classification tasks. There are no large fluctuations in the experimental results with changes in the hyper-parameter ω.

4) Triplet Sampling Methods: We consider the triplet sampling methods in our experiments because there exist $\mathcal{O}(N^3)$ candidate triplets in N training images. It is significant for the network to employ an effective triplet sampling method, which can contribute to rapid convergence and competitive performance. Given the anchor image, selecting the triplets randomly (*i.e.*, random sampling) during training always leads to slow convergence and suboptimal results. Adopting the hard sampling method may lead to poor training since there are mislabeled and poor quality images among the hardest samples. Inspired by [70], we utilize the semi-hard sampling method, in which every positive sample is taken into account and the semi-hard negative samples are randomly selected in a mini-batch. Moreover, the semi-hard sampling method results in $\mathcal{O}(N^2)$ triplets. More importantly, it makes the network converge quicker compared with the hard sampling, so we employ the semi-hard sampling in our experiments.

5) Ablation Study: To obtain thorough insight into our architecture, we further examine the advantage of each component through ablation experiments on the FI dataset. In Table II, the retrieval and classification results obtained using the feature representations extracted by models trained with different loss functions are presented. Compared with softmax loss, the metric learning constraints (*e.g.*, triplet) perform better in the affective image retrieval task. In contrast, their classification performance is inferior to the model with only the supervision of softmax loss, which demonstrates the superiority of softmax loss in discrimination of different classes. The performance on both tasks is obviously improved by integrating the softmax loss and the triplet loss in the training process. It is worth mentioning that the semi-hard sampling shows superiority compared with random sampling and hard sampling.

The model with the joint supervision of softmax loss and our sentiment similarity loss further improves the performance of retrieval, especially the 10% improvement in mAP$_2$, which demonstrates that our proposed loss can effectively capture the polarity of emotion. Moreover, the improvement of polarity discrimination promotes the framework to generate more dis-

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.2020.3001527, IEEE Transactions on Multimedia

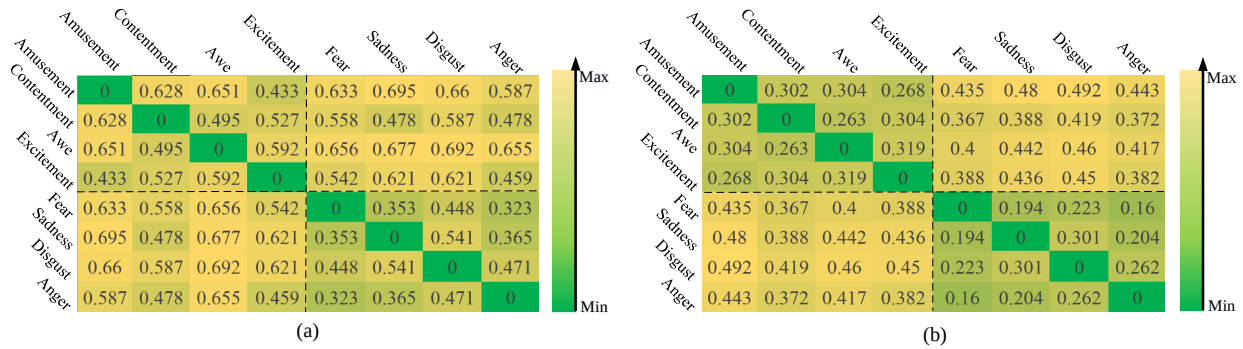


Fig. 8. Distances between the centroids of different class clusters in Euclidean space. The background of larger element in each matrix is greener, while the background of smaller element is yellower. (a) shows the results of the softmax loss, and (b) shows the results of our method.

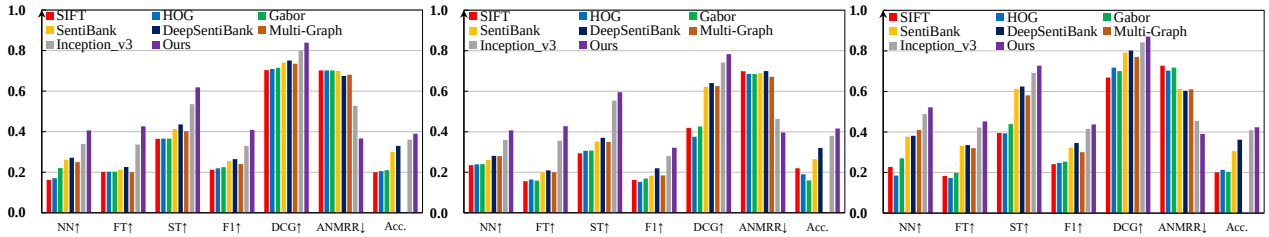


Fig. 9. Retrieval and classification results on Abstract, ArtPhoto, and IAPSa. We compare various baselines for learning the affective image representation, including conventional algorithms and CNN-based methods. Note that the multi-graph [11] method is only designed for affective image retrieval, thus the classification results are absent.

criminative features for specific emotion categories, resulting in better performance in terms of mAP_8. Further, the proposed sentiment vectors capture the rich texture information that is significant for image emotion representation. Therefore, compared with FC features, the sentiment vectors achieve better performance in both classification and retrieval tasks when employed to learn feature embedding. Last but not least, adaptively regulating the margin in the sentiment similarity loss further improves the results. In our method, we also find that using the semi-hard sampling method for the anchor, positive, related, and negative examples can obtain the optimal results, as shown in Table II.

To provide convincing and insightful results for affective image analysis, we visualize the distribution of features learned from different methods using t-SNE [89]. As shown in Figure 7, the features are mapped into a 2-D space. Each point represents a test image of FI (3,406 images in total), and the perplexity of t-SNE is set to 50. The feature embeddings learned from our framework are consistently better separated than the ones learned from the conventional softmax loss and combined loss consisting of softmax and triplet losses. It is observed that our method simultaneously enlarges the inter-class variance and shortens the intra-class distance, which benefits from incorporating emotion relations into the feature learning and adaptively adjusting the margin.

We also calculate the distance between different class clusters' centroids to verify that the classes of the same polarity are closer to each other in our learned embedding space. The mean of each feature embedding for the same class is regarded as the class cluster centroid. Each element of the symmetric matrices in Figure 8 represents the distance in Euclidean space between the centroids of different class clusters. Figure 8(a) shows the results calculated using the 2,048-dimensional features learned by softmax loss, while Figure 8(b) shows the results calculated using the 680-dimensional features of our method. In Figure 8(a), the distances between different emotional classes do not show a significant difference. The hierarchy of the label is not reflected. As can be seen in Figure 8(b), it is quite obvious that the distances between the emotional classes from the same polarity are smaller than those from different polarities. Therefore, our method is sensitive to sentiment polarity, which is essential to affective image retrieval.

To thoroughly verify the effectiveness of using the adaptive margin in our method, we conduct statistical significance analysis for the results obtained using the fixed margin and the adaptive margin, respectively. The experiments on mAP_8 are repeated five times as shown in Table III. Based on the two groups of data, we use one-way analysis of variance (one-way ANOVA) to conduct a significance test. The null hypothesis is that the difference between the two groups of results is not significant. The pre-specified significance level is set to 0.05, which indicates a 5% risk of concluding that a difference exists when there is no actual difference. The significance level of our test is p-value = $0.0039 < 0.05$. Therefore, the null hypothesis should be rejected. The results between using the fixed margin and adaptive margin show a significant difference.

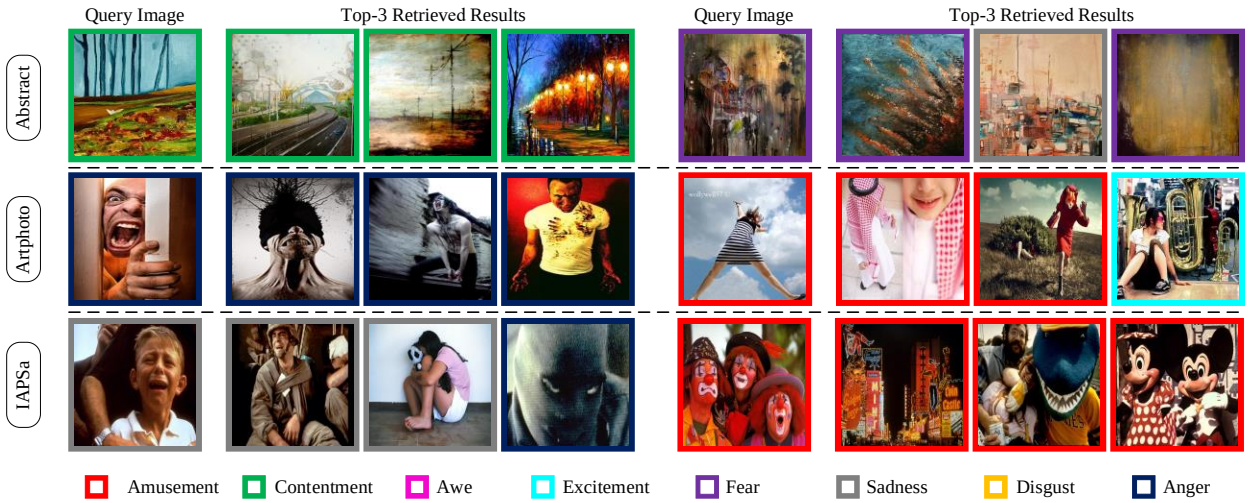


Fig. 10. Top-3 retrieval results of sampled query images from small-scale datasets based on the proposed method. The images in different color boxes belong to different emotion categories.

F. Results on Small-Scale Datasets

We also conduct comparison experiments with several other state-of-the-art methods on three widely used small-scale affective datasets, and the performances in retrieval and classification are shown in Figure 9. In the IAPSa and Abstract datasets, there are only 8 and 3 images in the 'anger' category, respectively. Therefore, we remove the images of the 'anger' category in the two datasets since they are not sufficient to perform 5-fold cross validation. The CNN-based methods outperform the conventional algorithm relying on hand-designed features in both retrieval and classification tasks due to the discriminative representation of deep features. Meanwhile, our framework achieves the best performance with the limited training samples, which demonstrates the generalization ability of the proposed method. Specifically, our method has a very substantial superiority to Inception-v3 pre-trained with softmax loss on the Abstract dataset consisting of color and texture. The sentiment vectors that can capture texture information considerably contribute to the substantial improvement. In Figure 10, we show the top-3 retrieval results for the query image from the three small datasets. The images belonging to 'Amusement' in ArtPhoto have an obvious gap with those in the FI dataset, which results in the failure case in the third place. Due to the smaller gap with the FI dataset, relatively better performance is achieved on IAPSa. Our method also shows favorable results on the Abstract dataset consisting of abstract paintings rather than natural photos, presenting a robust performance.

V. Discussion and Future Work

In this paper, we incorporate the hierarchy of emotion labels into the deep metric learning method and develop a unified multi-task framework that is jointly optimized by softmax and adaptive sentiment similarity losses. Specifically, we design the similarity loss considering the polarity of emotion. The margin in the similarity loss can adaptively adjusted under the guidance of the confidence score from the softmax layer to apply a stronger penalization on image pairs that are difficult to separate. To extract an informative texture representation, we design the sentiment vector as feature embedding, which is generated from the Gram matrices in multi-layers. Extensive experiments demonstrate that the proposed method performs favorably against the state-of-the-art algorithms on four widely used affective datasets.

An image can evoke multiple emotions. To rank the retrieved images more reasonably, it is significant to measure the emotional similarity based on the emotion distribution. In addition, the emotional stimuli in a small region may determine the dominant emotion of one image. Therefore, determining how to take local information into consideration for affective image retrieval and classification is also worth studying. Due to the limitation of the available datasets, we can only study affective image retrieval based on the datasets built for the classification task. In general, these datasets are coarse (*i.e.*, eight categories in Mikel's wheel) for retrieval, so the intra-class variation is always large. Therefore, it is worth constructing a dataset with fine-grained categories for the affective image retrieval task. In the future, it is likely that affective image retrieval can be applied to some specific domains, such as interior decoration, fashion analysis, and product design.

Acknowledgement

This work was supported by the Major Project for New Generation of AI Grant (NO.2018AAA0100403), NSFC (NO.61876094, U1933114), Natural Science Foundation of Tianjin, China (NO.18JCYBJC15400, 18ZXZNGX00110), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR), and the Fundamental Research Funds for the Central Universities. .

References

[1] J. Yang, D. She, Y.-K. Lai, and M.-H. Yang, "Retrieving and classifying affective images via deep metric learning," in *AAAI*, 2018.

[2] B. H. Detenber, R. F. Simons, and G. G. Bennett Jr., "Roll 'em!: The effects of picture motion on emotional responses," *J. Broadcast. & Electr. Media*, vol. 42, no. 1, pp. 113–127, 1998.

[3] W. Wang and Q. He, "A survey on emotional semantic image retrieval," in *ICIP*, 2008.

[4] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "RAPID: Rating pictorial aesthetics using deep learning," in *ACM MM*, 2014.

[5] S. Qian, T. Zhang, and C. Xu, "Multi-modal multi-view topic-opinion mining for social event analysis," in *ACM MM*, 2016.

[6] S. Zhao, H. Yao, Y. Gao, R. Ji, W. Xie, X. Jiang, and T.-S. Chua, "Predicting personalized emotion perceptions of social images," in *ACM MM*, 2016.

[7] B. Wu, J. Jia, Y. Yang, P. Zhao, J. Tang, and Q. Tian, "Inferring emotional tags from social images with user demographics," *IEEE Transactions on Multimedia*, vol. 19, no. 7, pp. 1670–1684, 2017.

[8] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *ACM MM*, 2010.

[9] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun, "Exploring principles-of-art features for image emotion recognition," in *ACM MM*, 2014.

[10] H. Zhang, Z. Yang, M. Gönen, M. Koskela, J. Laaksonen, T. Honkela, and E. Oja, "Affective abstract image classification and retrieval using multiple kernel learning," in *ICONIP*, 2013.

[11] S. Zhao, H. Yao, Y. Yang, and Y. Zhang, "Affective image retrieval via multi-graph learning," in *ACM MM*, 2014.

[12] A. Sartori, D. Culibrk, Y. Yan, and N. Sebe, "Who's afraid of itten: Using the art theory of color combination to analyze emotions in abstract paintings," in *ACM MM*, 2015.

[13] X. Lu, P. Suryanarayan, R. B. Adams Jr, J. Li, M. G. Newman, and J. Z. Wang, "On shape and the computability of emotions," in *ACM MM*, 2012.

[14] 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 *AAAI*, 2016.

[15] J. Yang, D. She, and M. Sun, "Joint image emotion classification and distribution learning via deep convolutional neural network," in *IJCAI*, 2017.

[16] 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 Res. Methods*, vol. 37, no. 4, pp. 626–630, 2005.

[17] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *AAAI*, 2015.

[18] C. Xu, S. Cetintas, K.-C. Lee, and L.-J. Li, "Visual sentiment prediction with deep convolutional neural networks," *arXiv:1411.5731*, 2014.

[19] C. Deng, E. Yang, T. Liu, J. Li, W. Liu, and D. Tao, "Unsupervised semantic-preserving adversarial hashing for image search," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 4032–4044, 2019.

[20] T. Rao, M. Xu, and D. Xu, "Learning multi-level deep representations for image emotion classification," *arXiv:1611.07145*, 2016.

[21] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *SIMBAD*, 2015.

[22] L. A. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," *Febs Letters*, vol. 70, no. 1, pp. 51–55, 2015.

[23] 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.

[24] S. Zhao, G. Ding, Q. Huang, T.-S. Chua, B. W. Schuller, and K. Keutzer, "Affective image content analysis: A comprehensive survey." in *IJCAI*, 2018.

[25] S. Fan, Z. Shen, M. Jiang, B. L. Koenig, J. Xu, M. S. Kankanhalli, and Q. Zhao, "Emotional attention: A study of image sentiment and visual attention," in *CVPR*, 2018.

[26] Y. Duan, W. Zheng, X. Lin, J. Lu, and J. Zhou, "Deep adversarial metric learning," in *CVPR*, 2018.

[27] W. Ge, "Deep metric learning with hierarchical triplet loss," in *ECCV*, 2018.

[28] M. A. Nicolaou, H. Gunes, and M. Pantic, "A multi-layer hybrid framework for dimensional emotion classification," in *ACM MM*, 2011.

[29] M. Solli and R. Lenz, "Color based bags-of-emotions," in *CAIP*, 2009.

[30] 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, 2017.

[31] H.-R. Kim, Y.-S. Kim, S. J. Kim, and I.-K. Lee, "Building emotional machines: Recognizing image emotions through deep neural networks," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 2980–2992, 2018.

[32] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[33] S. Zhao, X. Zhao, G. Ding, and K. Keutzer, "Emotiongan: unsupervised domain adaptation for learning discrete probability distributions of image emotions," in *ACM MM*, 2018.

[34] R. Panda, J. Zhang, H. Li, J.-Y. Lee, X. Lu, and A. K. Roy-Chowdhury, "Contemplating visual emotions: understanding and overcoming dataset bias," in *ECCV*, 2018.

[35] H. Xiong, H. Liu, B. Zhong, and Y. Fu, "Structured and sparse annotations for image emotion distribution learning," in *AAAI*, 2019.

[36] T. Chen, F. X. Yu, J. Chen, Y. Cui, Y.-Y. Chen, and S.-F. Chang, "Object-based visual sentiment concept analysis and application," in *ACM MM*, 2014.

[37] S. Zhao, G. Ding, Y. Gao, and J. Han, "Approximating discrete probability distribution of image emotions by multi-modal features fusion," in *IJCAI*, 2017.

[38] 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 *ACM MM*, 2013.

[39] V. Yanulevskaya, J. Van Gemert, K. Roth, A.-K. Herbold, N. Sebe, and J.-M. Geusebroek, "Emotional valence categorization using holistic image features," in *ICIP*, 2008.

[40] 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 *ACM MM*, 2012.

[41] X. Wang, J. Jia, J. Yin, and L. Cai, "Interpretable aesthetic features for affective image classification," in *ICIP*, 2013.

[42] J. Yuan, S. Mcdonough, Q. You, and J. Luo, "Sentribute: image sentiment analysis from a mid-level perspective," in *WISDOM*, 2013.

[43] C. Deng, Y. Xue, X. Liu, C. Li, and D. Tao, "Active transfer learning network: A unified deep joint spectral–spatial feature learning model for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 3, pp. 1741–1754, 2019.

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

[45] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009.

[46] 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.

[47] E. Ragusa, E. Cambria, R. Zunino, and P. Gastaldo, "A survey on deep learning in image polarity detection: Balancing generalization performances and computational costs," *Electronics*, vol. 8, no. 7, p. 783, 2019.

[48] S. Zhao, C. Lin, P. Xu, S. Zhao, Y. Guo, R. Krishna, G. Ding, and K. Keutzer, "Cycleemotiongan: Emotional semantic consistency preserved cyclegan for adapting image emotions," in *AAAI*, 2019.

[49] Y. He and G. Ding, "Deep transfer learning for image emotion analysis: Reducing marginal and joint distribution discrepancies together," *Neural Processing Letters*, pp. 1–10, 2019.

[50] J. Yang, D. She, Y.-K. Lai, P. L. Rosin, and M.-H. Yang, "Weakly supervised coupled networks for visual sentiment analysis," in *CVPR*, 2018.

[51] 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.

[52] S. Fan, M. Jiang, Z. Shen, B. L. Koenig, M. S. Kankanhalli, and Q. Zhao, "The role of visual attention in sentiment prediction," in *ACM MM*, 2017.

[53] C. Deng, X. Liu, C. Li, and D. Tao, "Active multi-kernel domain adaptation for hyperspectral image classification," *Pattern Recognition*, vol. 77, pp. 306–315, 2018.

[54] A. Bellet, A. Habrard, and M. Sebban, "A survey on metric learning for feature vectors and structured data," *arXiv:1306.6709*, 2013.

[55] E. Yang, C. Deng, C. Li, W. Liu, J. Li, and D. Tao, "Shared predictive cross-modal deep quantization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 11, pp. 5292–5303, 2018.

[56] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *CVPR*, 2014.

[57] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *CVPR*, 2014.

[58] S. Bai, X. Bai, Q. Tian, and L. J. Latecki, "Regularized diffusion process for visual retrieval," in *AAAI*, 2017.

[59] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *CVPR*, 2017.
[60] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3492–3506, 2017.
[61] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.
[62] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *CVPR*, 2005.
[63] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large scale online learning of image similarity through ranking," *Journal of Machine Learning Research*, vol. 11, pp. 1109–1135, 2010.
[64] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *NIPS*, 2016.
[65] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *CVPR*, 2016.
[66] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, 2016.
[67] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, "Triplet-center loss for multi-view 3d object retrieval," in *CVPR*, 2018.
[68] S. Kim, M. Seo, I. Laptev, M. Cho, and S. Kwak, "Deep metric learning beyond binary supervision," in *CVPR*, 2019.
[69] Y. Cui, F. Zhou, Y. Lin, and S. Belongie, "Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop," in *CVPR*, 2016.
[70] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015.
[71] B. Harwood, B. Kumar, G. Carneiro, I. Reid, T. Drummond *et al.*, "Smart mining for deep metric learning," in *ICCV*, 2017.
[72] W. Zheng, Z. Chen, J. Lu, and J. Zhou, "Hardness-aware deep metric learning," in *CVPR*, 2019.
[73] X. Zhu, L. Li, W. Zhang, T. Rao, M. Xu, Q. Huang, and D. Xu, "Dependency exploitation: A unified cnn-rnn approach for visual emotion recognition." in *IJCAI*, 2017.
[74] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Technical manual and affective ratings," *NIMH Center for the Study of Emotion and Attention*, vol. 1, pp. 39–58, 1997.
[75] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016.
[76] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.
[77] D. G. Lowe *et al.*, "Object recognition from local scale-invariant features." in *ICCV*, 1999.
[78] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.
[79] M. Idrissa and M. Acheroy, "Texture classification using gabor filters," *Pattern Recognition Letters*, vol. 23, no. 9, pp. 1095–1102, 2002.
[80] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski, "Orb: An efficient alternative to sift or surf." in *ICCV*, 2011.
[81] B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara, and S.-F. Chang, "Visual affect around the world: A large-scale multilingual visual sentiment ontology," in *ACM MM*, 2015.
[82] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.
[83] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *arXiv:1409.1556*, 2014.
[84] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *ECCV*, 2014.
[85] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, "End-to-end learning of deep visual representations for image retrieval," *IJCV*, vol. 124, no. 2, pp. 237–254, 2017.
[86] D. Yi, Z. Lei, and S. Li, "Deep metric learning for practical person re-identification," in *ICPR*, 2014.
[87] Q. Qian, L. Shang, B. Sun, J. Hu, H. Li, and R. Jin, "Softtriple loss: Deep metric learning without triplet sampling," in *ICCV*, 2019.
[88] F. Cakir, K. He, X. Xia, B. Kulis, and S. Sclaroff, "Deep metric learning to rank," in *CVPR*, 2019.
[89] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.