Attention-Aware Polarity Sensitive Embedding for Affective Image Retrieval

Xingxu Yao[1], Dongyu She[1], Sicheng Zhao[2]✉, Jie Liang[1], Yu-Kun Lai[3], Jufeng Yang[1]✉
[1]Nankai University [2]University of California, Berkeley [3]Cardiff University
{yxx_hbgd, sherry6656}@163.com, schzhao@gmail.com, liang27jie@163.com
LaiY4@cardiff.ac.uk, yangjufeng@nankai.edu.cn

Abstract

Images play a crucial role for people to express their opinions online due to the increasing popularity of social networks. While an affective image retrieval system is useful for obtaining visual contents with desired emotions from a massive repository, the abstract and subjective characteristics make the task challenging. To address the problem, this paper introduces an Attention-aware Polarity Sensitive Embedding (APSE) network to learn affective representations in an end-to-end manner. First, to automatically discover and model the informative regions of interest, we develop a hierarchical attention mechanism, in which both polarity- and emotion-specific attended representations are aggregated for discriminative feature embedding. Second, we present a weighted emotion-pair loss to take the inter- and intra-polarity relationships of the emotional labels into consideration. Guided by attention module, we weight the sample pairs adaptively which further improves the performance of feature embedding. Extensive experiments on four popular benchmark datasets show that the proposed method performs favorably against the state-of-the-art approaches.

1. Introduction

With the increasing popularity of online social networks, people are more likely to express their opinions through posting images on social platforms such as Flickr and Instagram. Recently, affective image analysis that studies the emotional response of humans on visual stimuli has drawn attention from both psychologists [38, 49, 32] and computer vision researchers [30, 63] due to its wide applicability, *e.g.* opinion mining [36, 39], image captioning [8, 31], *etc.*

How to search affective images based on human perception is a meaningful yet challenging task. Various emotion-based image retrieval (EBIR) systems have been proposed [54, 24, 34, 65]. Compared to content-based image retrieval (CBIR), EBIR involves high-level abstract semantics and human perception subjectivity. To bridge the "affective gap" between low-level features and high-level affective semantics, some hand-crafted features are proposed according to psychology and art theory [30, 68]. To capture the semantic similarity among affective images, Zhao *et al.* [72] employ multi-graph learning for affective image retrieval based on features of different levels including low-level color, texture, and other high-level features that contribute to expressing image emotions. More recently, deep learning has been harnessed to predict emotions evoked by images via embedding images into a feature space [58, 41, 52, 47], which results in breakthrough performance. Pang *et al.* [37] develop a unit density model over the multi-modal space using a deep Boltzmann machine, which enables emotion-oriented cross-modal retrieval. Yang *et al.* [57] propose a multi-task framework to simultaneously optimize the classification and retrieval losses, in which the performances of both tasks are boosted.

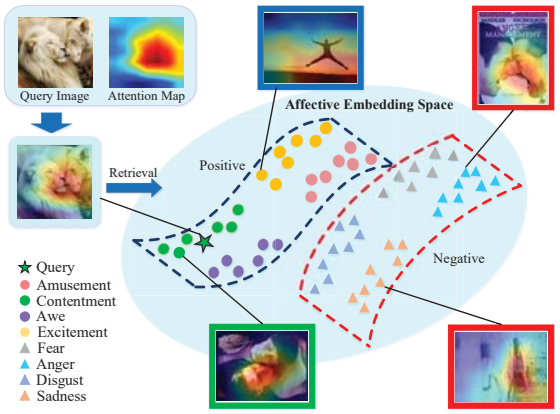


Figure 1. Illustration of retrieving affective images in the embedding space. The two regions in the space represent binary sentiment polarities, *i.e.* positive and negative. For the given query image, the retrieved image from exactly the same emotion category is shown in a green box, while the images from the same polarity but different category and the opposite polarity are in blue and red boxes, respectively.

However, there are two important characteristics in visual emotion (shown in Fig. 1), which are neglected in existing methods for affective image retrieval. On the one hand, informative regions of interest are crucial to image emotion (see the heat map of each sample image) [12, 3, 50], which can evoke emotional stimuli to people; on the other hand, there exist sentiment polarities in emotional label space other than concrete categories. Note that polarity indicates the coarse-level classes {positive, negative}, and the concrete-level emotions are defined as {amusement, contentment, awe, excitement, fear, anger, disgust, sadness} as per [32, 63]. In this paper, the term 'class' is utilized to mean both sentiment polarity and emotion category. Given a query image, our goal is to rank the retrieved images according to the relationship with the given image in the following order: the same emotion category, the same polarity but different emotion categories, different polarity.

In the paper, we propose an attention-aware polarity sensitive embedding (APSE) network for affective image retrieval according to aforementioned characteristics of visual emotion. In detail, there exists a correlation between sentiment polarity and low-level features [42, 29, 68], while specific emotion categories are mainly determined by semantic content. Therefore, in the attention module, we utilize the *polarity-specific* attention in lower layers of the network, and exploit *emotion-specific* attention in higher layers. In the embedding process, we introduce a polarity sensitive feature embedding strategy based on the proposed weighted emotion-pair (WEP) loss. We separate binary sentiment polarities in the embedding space, while also effectively distinguishing different emotions in the same polarity. Guided by the attention module, hard negative examples are imposed stronger penalty so as to improve the learning performance. The unified architecture is optimized by the total loss consisting of WEP and attention losses to learn discriminative feature embedding.

Our contributions are twofold. 1) We propose to take multi-level attended local features into account for affective image retrieval, based on the observation that low-level and high-level image features concern different levels of the emotion hierarchy. 2) We introduce an attention-aware polarity sensitive embedding (APSE) network, which takes the inter- and intra-polarity relationships of the emotional labels into consideration. Our proposed WEP loss effectively connects the attention module and embedding process for more effective learning. Extensive experiments demonstrate the effectiveness of the proposed method.

2. Related Work

2.1. Visual Emotion Analysis

In the field of visual emotion analysis, most existing methods focus on emotion prediction [73, 35, 70, 57, 62, 69, 40, 23]. Early work uses a variety of hand-crafted features [30, 60] including shape features [29] and principles-of-art features [68] to represent the emotions evoked by images. In addition, Borth *et al*. [2] propose adjective noun pairs (ANP) to bridge the affective gap between low-level features and high-level emotion semantics. With extensive applications of deep learning models, numerous methods [52, 41, 74] exploit convolutional neutral networks (CNNs) to extract deep features for emotion representations, which perform well on image emotion classification [6, 56, 58], emotion label distribution prediction [71, 67], and affective image retrieval [72].

While many methods have been devoted to image emotion prediction, far less attention is paid to affective image retrieval. Wang *et al*. [54] propose an EBIR system that allows users to perform retrieval using sentiment semantic words, and the system is further improved for different tasks [24, 34]. Zhao *et al*. [72] utilize multi-graph learning to retrieve affective images that are similar to the query image in emotion. A deep framework which simultaneously optimizes the classification and retrieval tasks is proposed in [57]. Different from the existing methods, we develop a polarity sensitive embedding method based on multi-level attended features for affective image retrieval.

2.2. Visual Attention Mechanism

Attention mechanism is widely used in various visual tasks [44, 55, 1, 66, 5, 4, 11], since it can find image regions that play a decisive role in networks. Wang *et al*. [51] train deep residual networks for image classification by introducing an attention based learning method. SCA-CNN network integrating spatial and channel-wise attention is proposed in [4] for image captioning. According to psychological theory [50, 12], affective content is easier to hold human attention than non-affective content. Unlike specific salient objects which have well-defined boundaries, the region arousing emotion may be ambiguous and abstract [56].

For affective images, prior methods [59, 61] detect emotional attention regions from numerous candidate bounding boxes, increasing the computational burden. Our method generates soft attention maps with the single shot based on the feature activations in an end-to-end manner. Moreover, we integrate features from multiple layers and build a hierarchical attention mechanism for learning robust representations in the embedding space. That is, both polarity-specific features from lower layers and emotion-specific features from higher layers are combined together in our framework.

2.3. Feature Embedding Learning

Recently, numerous methods have utilized embedding learning to measure image similarity for various tasks [28, 9, 17, 64, 53, 20]. Based on the popular pairwise loss [10],

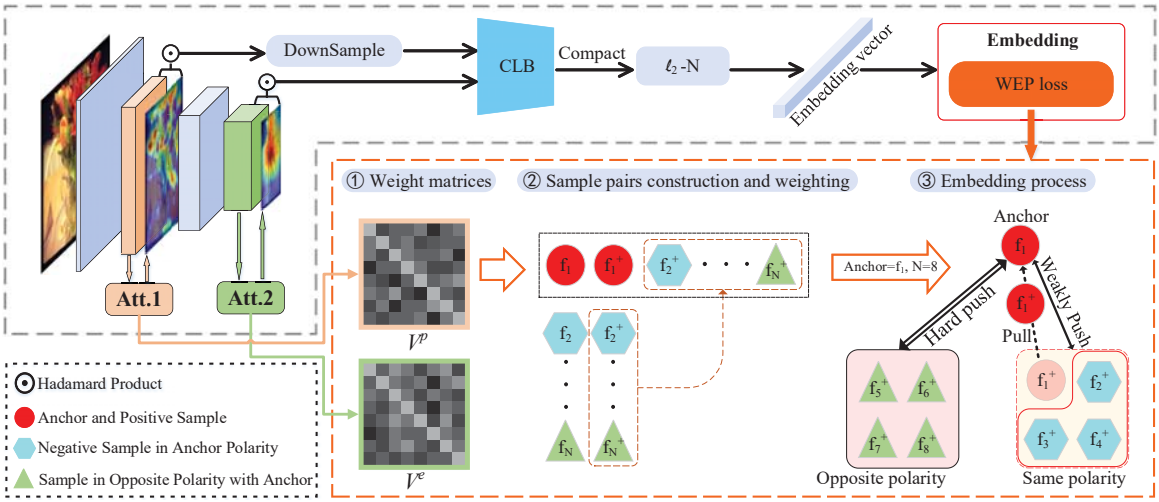


Figure 2. Pipeline of the proposed approach. In the weighted emotion-pair (WEP) loss, we employ similar emotion categories in the FI dataset [63] with the number of categories $N = 8$. Here, four categories are positive and the other four are negative. The details of the process of generating attention maps are presented in Fig. 3. Att.1 and Att.2 represent polarity-specific attention and emotion-specific attention, respectively. V^p and V^e mean polarity and emotion-level weight matrices for sample pairs. CLB denotes cross-level bilinear operation. f_i and f_i^+ represent the features of anchor point and positive example from the i^{th} category, respectively.

Song *et al*. [33] utilize a matrix consisting of pairwise distances of the mini-batch to create a loss function which incorporates all samples to form a lifted embedding structure. In order to produce effective training samples, Harwood *et al*. [18] conduct a smart mining procedure to train the model effectively. In addition, Duan *et al*. [15] employ deep adversarial learning to generate hard negatives from easy negatives for building more robust models. Motivated by the fact that emotional classes have a hierarchical relationship, *i.e*., from coarse polarity to concrete emotions, we develop polarity-sensitive WEP loss to measure the similarity of the query and the retrieved images.

3. Methodology

We propose APSE network which can be trained in an end-to-end manner. It contains two main closely related components, as shown in Fig. 2. First, the proposed method integrates polarity- and emotion-specific attended features extracted by hierarchical attention mechanism (Sec. 3.1). Second, we learn polarity-sensitive and discriminative feature embedding by optimizing WEP loss guided by the attention module (Sec. 3.2).

3.1. Hierarchical Attention Mechanism

In addition to the regions for specific emotions obtained from higher layers in the deep network, we also learn the attended regions for specific polarities from lower layers. We propose a simple yet effective attention mechanism (Fig. 3), whose module consists of attention head and output head, which is applied to both attention levels.

The attention head receives the l^{th} level feature activations $F^l \in \mathbb{R}^{c\times h\times w}$ as input, and outputs K^l attention maps, where c, h and w are the number of channels, and the height and width of the feature activations, and K^l represents the number of corresponding labels for layers at the l^{th} level. First, we sum up the received feature activation tensor through the channel direction. Thus, an $h \times w$ 2-D aggregation map A^l is derived from 3-D feature activations F^l, *i.e*., $A^l = \sum_{n=1}^{c} F_n^l$. Then a spatial attention mask Z^l is obtained by spatial-wise softmax operation on A^l. Based on Z^l, we implement spatial-wise attention on the feature activations F^l resulting in spatially-attended feature maps, *i.e*., $\hat{F}^l = F^l \odot Z^l$, where $\odot$ denotes Hadamard Product by broadcasting, *i.e*. repeating Z^l for each channel of F^l. Then a 1×1 conv layer is applied to reduce the dimension of $\hat{F}^l$ to $K^l \times h \times w$, denoted as $S^l \in \mathbb{R}^{K^l\times h\times w}$, with each 2-D feature activation corresponding to a sentiment polarity or specific emotion category depending on the level. S^l is put through a global average pooling layer and a softmax layer successively, resulting in confidence score vector $\boldsymbol{C}^l$ whose elements lie in the range of $[0, 1]$ and sum to 1.

The output head at the l^{th} level receives 2-D feature activations S^l and corresponding confidence scores. Each confidence score $\boldsymbol{c}$ can be regarded as the degree of tendency towards the corresponding class. Therefore, final attention map $\mathcal{U}$ is obtained by adding up all 2-D feature activations S_j weighted by confidence scores:

$$\mathcal{U} = norm(\sum_{j=1}^{K} \boldsymbol{c}_j S_j), \tag{1}$$

where $norm$ denotes the normalization operation. Note that

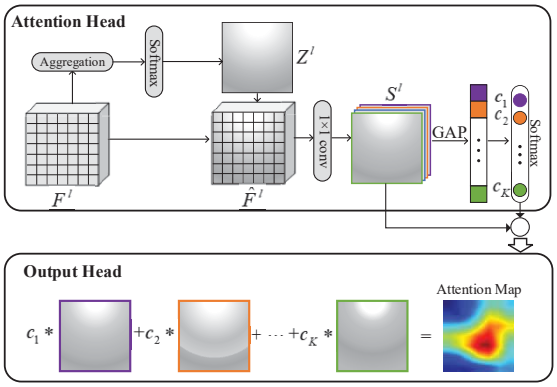


Figure 3. Overview of attention map generation. The class-aware activation and corresponding confidence score are derived in the attention head. In the output head, the attention map is obtained by weighting each activation map. In the lower layers, the attention module generates a polarity-specific attention map, whereas an emotion-specific attention map is generated in higher layers.

$K = 2$ in lower layers means binary sentiment polarities, while $K = 8$ in higher layers denotes eight emotion categories in Mikels' wheel [32]. Afterward, $\mathcal{U}$ is element-wise multiplied with $\hat{F}$ by broadcasting so as to generate discriminative attended features $F_w = \hat{F} \odot \mathcal{U}$. Based on labels of different hierarchies, we can assign different constraints in the same form on layers of different depths. Therefore, the attention loss can be drawn with the following unified formula:

$$\mathcal{L}_{att} = -\frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{K}\mathbf{1}[z_m = j]\log \boldsymbol{c}_j, \tag{2}$$

where $\mathbf{1}[s] = 1$ if the condition s is true, and 0 otherwise. M denotes the number of input images, and z_m the corresponding label ID of the m^{th} input image $\mathcal{I}_m$. The attention loss is exploited in both lower and higher layers simultaneously. What is different is that lower layers are supervised by binary polarities, while higher layers are supervised by eight specific emotion categories.

The features from different layers put particular emphasis on different information [41, 52, 74]. For the sake of integrating polarity- and emotion-specific attended features effectively, we use the bilinear operation [26] to make them interact with each other. We first downsample attended feature activation output from the low-level layer to the size of the attended feature activations from the high-level layer. Then we utilize the cross-level bilinear operation (CLB) to model the interactions of features of different levels and establish pairwise correlations between the channels.

3.2. Polarity Sensitive Embedding Learning

In this section, considering the polarity characteristic of sentiment, we propose the polarity sensitive emotion-pair (EP) loss inspired by N-pair loss. In the embedding process, sample pairs are further adaptively weighted based on confidence scores from the attention module, generating WEP loss. Specifically, the harder anchor-negative pairs are to separate, the higher the weight of them should be, so as to augment their proportion when training the network.

Review on N-pair loss. Given N categories, the N-pair loss function proposed in [46] optimizes to identify a positive example from $N-1$ negative examples. Define $\{(f_1, f_1^+), \cdots, (f_N, f_N^+)\}$ as N pairs of convolution features from N different categories, where f_i denotes the i^{th} category anchor point, and f_i^+ represents a positive example of the i^{th} category. Meanwhile, f_i^+ can also be regarded as a negative example of the j^{th} category ($\forall i \neq j$). The value of $f^\top f^+$ has positive correlation with the similarity between f and f^+. Therefore, the N-pair loss function can be formulated as

$$\mathcal{L}_{np} = \frac{1}{N}\sum_{i=1}^{N}\log(1 + \sum_{i\neq j}\exp(f_i^\top f_j^+ - f_i^\top f_i^+)). \tag{3}$$

EP loss. In general, N-pair loss can embed features effectively and efficiently. However, for affective image retrieval, the polarity characteristic cannot be considered by the approach directly. Therefore, it is essential to differentiate different negative examples based on their polarity when learning feature embedding. More specifically, image features from the same polarity should be more similar than those from opposite polarity. Therefore, our proposed inter-polarity loss is formulated as

$$\begin{aligned}\mathcal{L}_{inter} = \frac{1}{N}\sum_{i=1}^{N}\log(1 + \exp(&\frac{1}{N_{\mathcal{Q}_i}}\sum_{j\in\mathcal{Q}_i} f_i^\top f_j^+ \\ &- \frac{1}{N_{\mathcal{P}_i}}\sum_{j\in\mathcal{P}_i, i\neq j} f_i^\top f_j^+)),\end{aligned} \tag{4}$$

where $\mathcal{P}_i$ and $\mathcal{Q}_i$ denote the sets of emotion categories with the same and opposite polarities to the anchor from the i^{th} category, respectively. $N_{\mathcal{P}_i}$ and $N_{\mathcal{Q}_i}$ are the numbers of the corresponding categories.

The inter-polarity loss is very important for affective image retrieval, because it can largely avoid dramatic failure that the retrieved result has many images with opposite sentiment polarity, which may cause unpleasant user experience. That is, the inter-polarity loss ensures the returned images are consistent with query images in sentiment polarity. Further, the more challenging task is to learn discriminative feature embedding within the same polarity. To achieve this, we introduce a new intra-polarity loss to differentiate similar categories in the same polarity as follows:

$$\mathcal{L}_{intra} = \frac{1}{N}\sum_{i=1}^{N}\log(1 + \sum_{j\in\mathcal{P}_i, i\neq j}\exp(f_i^\top f_j^+ - f_i^\top f_i^+)). \tag{5}$$

Therefore, the EP loss is obtained by combining inter-polarity loss and intra-polarity loss as:

$$\mathcal{L}_{ep} = \mathcal{L}_{inter} + \mathcal{L}_{intra}. \tag{6}$$

Weighting sample pairs. Given an affective image $\mathcal{I}$, we can obtain its confidence scores regarding both polarity and emotion as demonstrated in Sec. 3.1. For an anchor $\mathcal{I}_i^a$ from the i^{th} category and one of its negative samples $\mathcal{I}_j^n$ from the j^{th} category, a higher confidence of $\mathcal{I}_i^a$ w.r.t. the j^{th} category or $\mathcal{I}_j^n$ w.r.t. the i^{th} category denotes that the pair is harder to separate. Consequently, we assign a stronger penalty term on this pair in the training process.

Specifically, $\boldsymbol{c}_{ij}^p$ and $\boldsymbol{c}_{ij}^e$ represent the confidences of $\mathcal{I}_i^a$ w.r.t. the j^{th} category in polarity- and emotion-level, while $\boldsymbol{c}_{ij}^{+p}$ and $\boldsymbol{c}_{ij}^{+e}$ represent the confidences of $\mathcal{I}_i^n$ w.r.t. the j^{th} category in polarity- and emotion-level. The weights are formed as

$$v_{ij}^p = \exp(\boldsymbol{c}_{ij}^p) \cdot \exp(\boldsymbol{c}_{ji}^{+p}), \tag{7}$$

$$v_{ij}^e = \exp(\boldsymbol{c}_{ij}^e) \cdot \exp(\boldsymbol{c}_{ji}^{+e}), \tag{8}$$

where v_{ij}^p denotes the polarity-level weight of the pair constructed by $\mathcal{I}_i^a$ and $\mathcal{I}_j^n$, and v_{ij}^e denotes the emotion-level weight of the pair constructed by $\mathcal{I}_i^a$ and $\mathcal{I}_j^n$. Note that v_{ij}^p will be set to be 1 if the i^{th} and j^{th} categories belong to the same polarity. Then v_{ij}^p and v_{ij}^e form the weight matrix V^p and V^e respectively as shown in Fig. 2, whose diagonal elements are set to 1 (*i.e.* $v_{ii}^p = 1$, $v_{ii}^e = 1$). The final weight $\widetilde{v}_{ij} = v_{ij}^p \cdot v_{ij}^e$. The value of $\widetilde{v}_{ij}$ $(\forall i \neq j)$ determines the importance during learning. We set the weight of any anchor-positive pair to be 1, *i.e.*, $\widetilde{v}_{ii} = 1$. Therefore, we introduce WEP (weighted EP) loss:

$$\begin{aligned}\mathcal{L}_{wep} = \frac{1}{N}\sum_{i=1}^{N}\log[(1+\exp(\frac{1}{N_{\mathcal{Q}_i}}\sum_{j\in\mathcal{Q}_i}\widetilde{v}_{ij}f_i^\top f_j^+ \\ - \frac{1}{N_{\mathcal{P}_i}}\sum_{j\in\mathcal{P}_i, i\neq j}\widetilde{v}_{ij}f_i^\top f_j^+))(1+\sum_{j\in\mathcal{P}_i, i\neq j}\exp(\widetilde{v}_{ij}f_i^\top f_j^+ \\ - f_i^\top f_i^+))].\end{aligned} \tag{9}$$

We define the total loss consisting of attention and WEP losses to optimize the unified embedding network simultaneously:

$$\mathcal{L}_{total} = \lambda\mathcal{L}_{wep} + (1-\lambda)\mathcal{L}_{att}, \tag{10}$$

where λ is the weight to control the trade-off between two types of losses.

4. Experiments

In this section, we conduct extensive experiments on the most commonly used affective datasets to evaluate the proposed algorithm against the state-of-the-art methods.

4.1. Datasets

We perform our experiments on four popular datasets, including Flickr and Instagram (FI) [63], Subset A of IAPS (IAPSa) [32], Artistic dataset (ArtPhoto) [30], and Abstract paintings (Abstract) [30]. FI is collected from social websites by querying Mikels' eight emotions as keywords, resulting in 23,308 labeled images. IAPSa consists of 395 images collected from International Affective Picture System (IAPS) [32], while ArtPhoto contains 806 artistic photographs searched by emotion categories. The Abstract is composed of 228 peer rated abstract paintings which contain abundant color and texture.

4.2. Evaluation Metrics

Following previous work [72, 57], we adopt the following metrics as our evaluation criteria. The mean precision of the retrieval results are represented by mean Average Precision (mAP). We concern both mAP of eight emotion-specific categories (mAP_8) and mAP of the two polarities (mAP_2). Nearest neighbor rate (NN) denotes the proportion of the rank-1 sample belonging to the same category with the query. First tier (FT) and second tier (ST) both represent the recall of the retrieval results. FT denotes the top-$\boldsymbol{n}$ recall, while ST is defined as the top-$2\boldsymbol{n}$ recall. Here, $\boldsymbol{n}$ is the number of all the positive examples. Discounted cumulative gain (DCG) [21] measures the importance of different positions of relevant samples in the ranking sequence of returned results. F_1 score is a measure combining Precision and Recall, as their harmonious mean. Average normalized modified retrieval rank (ANMRR) [16] considers the ranking sequence of relevant images within the retrieved results. Smaller values of ANMRR represent better retrieval results, while for other evaluation metrics the larger the better.

4.3. Baselines

We compare the proposed method with different baselines. First, we extract low-level local descriptors (*i.e.* SIFT and HOG), whose dimensions are fixed to 1000. We also extract mid-level features, including 1200-dimensional ANP detectors of SentiBank [2], 2089-dimensional features of DeepSentiBank [7], and 4342-dimensional features of MVSO (English) [22]. For CNN methods, we fine-tune deep models with softmax loss based on different architectures, including AlexNet, VggNet, GoogleNet, and ResNet-50, and extract the features from the last FC layer for retrieval. Also, we train different embedding learning methods based on ResNet-50, including contrastive loss [10], triplet loss [43], N-pair loss [46], and retrieve images using 2048-dimensional features. We also compare with the state-of-the-art methods for affective image retrieval, including Yang *et al*. [57] and Multi-Graph [72].

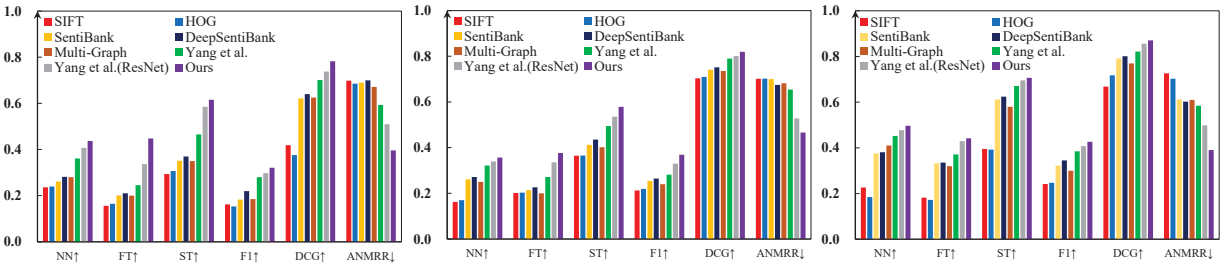
4.4. Implementation Details

Following [57], each image in the test set of FI dataset is treated as a query image to retrieve relevant images in the training set. For small-scale datasets, we use each image to retrieve the rest of images. We rank the retrieved images based on the similarity between them and the query image.

The proposed framework is based on ResNet-50 [19] pre-trained on the ImageNet [14]. The original images are

Table 1. Retrieval performance on the FI dataset. We evaluate the proposed method against different algorithms, including traditional methods (TRA), existing CNN models (CNN), and embedding learning methods (EMB). Note that 'S' denotes using softmax loss for training, and 'Dim.' represents the dimension of features.

	Methods	Dim.	mAP_8 ↑	mAP_2 ↑	FT↑	ST↑	NN↑	DCG↑	ANMRR↓
TRA	SIFT [27]	1000	0.1705	0.5913	0.1830	0.3513	0.2462	0.4507	0.6553
	HOG [13]	1000	0.2115	0.6002	0.1926	0.3620	0.3225	0.4639	0.6424
	Sentibank [2]	1200	0.2337	0.6168	0.2422	0.4232	0.3990	0.5223	0.5934
CNN	DeepSentiBank [7]	2089	0.2559	0.6247	0.2658	0.4468	0.4583	0.5509	0.5655
	MVSO [22]	4342	0.2798	0.6366	0.2877	0.4761	0.5158	0.5731	0.5346
	AlexNet (S) [25]	4096	0.2709	0.6328	0.2795	0.4693	0.5038	0.5633	0.5463
	VggNet (S) [45]	4096	0.3013	0.6552	0.3007	0.4887	0.5511	0.5860	0.5161
	GoogleNet (S) [48]	2048	0.3583	0.6773	0.3571	0.5619	0.5816	0.6403	0.4517
	ResNet (S) [19]	2048	0.4380	0.7068	0.4286	0.6079	0.6084	0.6816	0.3998
	WSCNet [56]	4096	0.5060	0.7381	0.4653	0.6223	0.6358	0.6910	0.3872
EMB	Contrastive loss (ResNet) [10]	2048	0.3842	0.6972	0.3768	0.5702	0.5711	0.6508	0.4396
	Triplet loss (ResNet) [43]	2048	0.5130	0.7120	0.4864	0.6216	0.5710	0.6843	0.3860
	N-pair loss (ResNet) [46]	2048	0.5217	0.8062	0.4785	0.7075	0.5341	0.7310	0.3089
	Yang *et al*. (GoogleNet) [57]	640	0.4885	0.8098	0.4834	0.6978	0.6023	0.7802	0.3135
	Yang *et al*. (ResNet) [57]	544	0.6395	0.8081	0.5995	0.7354	0.6164	0.7866	0.2518
	APSE (ours)	512	**0.7344**	**0.9079**	**0.6985**	**0.7817**	**0.6613**	**0.8114**	**0.2201**



(a) Artphoto (b) Abstract (c) IAPSa

Figure 4. Retrieval performance on the three small datasets (Artphoto, Abstract, and IAPSa).

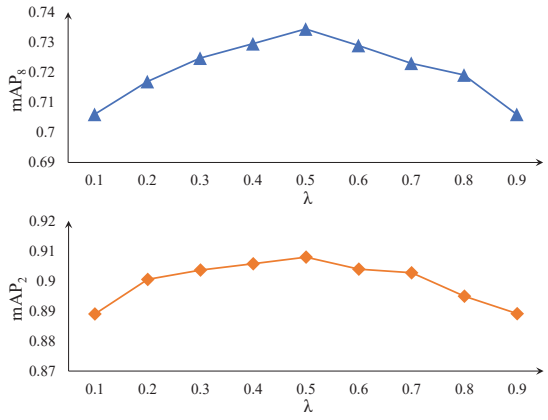


Figure 5. Effect of λ for total loss on mAP_8 and mAP_2 testing on FI dataset. Note that λ is the weight of $\mathcal{L}_{wep}$, and $1-\lambda$ is the weight of $\mathcal{L}_{att}$.

resized to 256 × 256 followed by a center 224 × 224 cropping. We initialize the learning rate as 0.001 and drop down one-tenth every 40 epochs. The gross number of epochs is 100 for fine tuning all layers by stochastic gradient descent (SGD) with a batch size of 32 ensuring images from each emotion. We optimize the parameters of the framework by SGD with the weight decay of 0.0005 and a momentum of 0.9. Considering both effectiveness and consumption of parameters, we choose the features from last layer of $conv_3$ and $conv_5$ to represent the low-level and high-level features, respectively. For contrastive and triplet losses, we set the margin γ to 0.4 and 0.2 respectively. We adopt the semi-hard triplet sampling method in triplet loss. In our architecture, the dimension of the output embedding feature after being compacted is 512 according to the experience from [26]. The FI dataset is split randomly into 80% training, 5% validation, and 15% testing sets. For small-scale datasets, we transfer the parameters of the network fine-tuned on FI to them. 5-fold validation is performed and the average performance is reported.

4.5. Retrieval Performance

We evaluate the retrieval performance with different methods on four affective datasets. As shown in Tab. 1,

Table 2. Ablation experiments on the FI dataset. The fundamental framework is ResNet-50 pre-trained on ImageNet. Here, AT represents the attention loss consisting of two softmax losses. HA denotes hierarchical attention, and SA denotes the emotion-specific attention on the last convolutional layer. CLB represents cross-level bilinear operation. SO means using the feature from the last convolution layer, and MO means using the feature from the last layer from both $conv_3$ and $conv_5$, respectively. When CLB is not selected, the features from different layers are concatenated directly. The weights of all parts in the combined loss are the same.

AT	N-pair	EP	WEP	SA	HA	CLB	SO	MO	mAP_8 ↑	mAP_2 ↑	FT↑	ST↑	NN↑	DCG↑	ANMRR↓
√							√		0.4380	0.7068	0.4286	0.6079	0.6084	0.6816	0.3998
	√						√		0.5217	0.8062	0.4785	0.7075	0.5341	0.7310	0.3089
		√					√		0.5680	0.8558	0.5247	0.7187	0.5623	0.7602	0.2789
√	√						√		0.6225	0.7816	0.5779	0.7255	0.5975	0.7451	0.2623
√		√					√		0.6430	0.8241	0.6036	0.7485	0.6110	0.7863	0.2551
√		√						√	0.6680	0.8325	0.6365	0.7504	0.6278	0.7885	0.2421
√		√		√				√	0.6938	0.8605	0.6417	0.7604	0.6290	0.7883	0.2396
√		√			√			√	0.7051	0.8733	0.6696	0.7595	0.6393	0.7952	0.2388
√		√			√	√		√	0.7190	0.8912	0.6824	0.7677	0.6495	0.8052	0.2294
√			√		√	√		√	**0.7344**	**0.9079**	**0.6985**	**0.7817**	**0.6613**	**0.8114**	**0.2201**

we compare our proposed method with traditional methods, CNN-based methods and other embedding learning methods on FI. We can see that current popular deep representations outperform the hand-crafted features. In general, embedding learning methods get remarkable improvements in all evaluation metrics other than NN as presented in Tab. 1, compared with the CNN architectures trained by softmax loss. This is because softmax loss only concerns the location of single data rather than a holistic distribution in metric space. In addition, we compare our method with other competitive and influential embedding learning approaches as well as the state-of-the-art algorithms. For fair comparison, we also implement the state-of-the-art [57] using ResNet-50 architecture as this work. Our framework improves about 10% on mAP_8 and mAP_2 respectively as compared to state of the arts. The other evaluation metrics are also improved obviously.

For other three small-scale datasets, we transfer the model trained on the FI dataset for fine-tuning on the target datasets. As reported in Fig. 4, we draw similar conclusions on the small-scale datasets as FI, where the proposed method still obtains the best retrieval results. This illustrates that our framework has robust generalization ability.

4.6. Influence of Parameter λ

In Eqn. (10), the value of λ controls the relative importance between the WEP loss and attention loss. The bigger the value of λ is, the more important the WEP loss is. We use the two essential metrics, which are mAP_8 and mAP_2, on FI dataset to demonstrate how λ influences the performance of total loss on FI. Note that the two losses are not isolated absolutely, so we only concern the results with λ ranging from 0.1 to 0.9. As shown in Fig. 5, we can find through the curves that: (1) mAP_8 is more sensitive than mAP_2 for the variation of λ; (2) When $\lambda = 0.5$, mAP_8 and mAP_2 both achieve the best performance. On the whole, the values of the two metrics are stable, which demonstrates that our method is robust for affective image retrieval.

4.7. Ablation Study

In order to demonstrate the contribution of different components in the proposed method, we further examine the advantage of each component through ablation experiments on FI dataset. First, AT is the attention loss consisting of two softmax losses on $conv_3$ and $conv_5$, respectively. As shown in the first part of Tab. 2, our EP loss has obvious superiority compared with the softmax and N-pair losses in all criteria. The results on mAP_8 and mAP_2 illustrate the architecture optimized by the EP loss improves the precision of retrieved images considering sentiment polarities other than specific emotions. As can be seen, integrating the AT and EP losses can enhance the performance on all the evaluation criteria other than mAP_2, because they benefit each other in the process of training. On the one hand, the AT provides category-specific cues for EP loss; on the other hand, the AT in the last convolution layer neglects the distinction between polarity, resulting in a weak decline on mAP_2, which can be recovered in our attention mechanism and multi-level output.

In addition, experiments are also performed to verify the effect of attention mechanism as shown in the second part of Tab. 2. The result of only using SA exceeds about 3% on both mAP_8 and mAP_2 compared with the performance of framework without any attention. Furthermore, hierarchical attention mechanism also has obvious benefits compared with SA, when both of them utilize features from both $conv_3$ and $conv_5$. It demonstrates that the attended features from different levels are complementary, resulting in improvement on overall retrieval performance.

In order to make the features from different levels interact effectively, the cross-level bilinear (CLB) is exploited to integrate multi-level information, leading to further per-

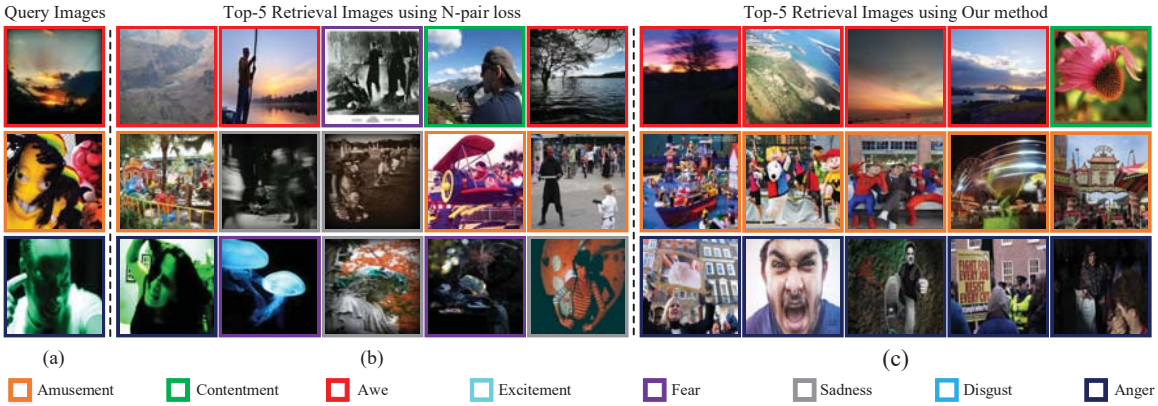


Figure 6. Top 5 results of sample query images from the FI dataset. (a) are sample query images from FI. (b-c) are the retrieval results of networks trained by the N-pair loss and our method, respectively. Image frames with different colors represent different emotions.

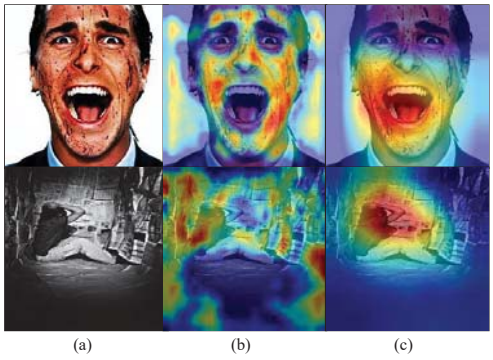


Figure 7. Visualization of attention maps from different levels. The images from the FI dataset are presented in column (a), and the visualizations of polarity- and emotion-specific attention results are presented in column (b) and column (c), respectively. The classes of the two sample images are disgust and sadness, respectively.

formance improvement over the baseline that directly fuses them by concatenation. More importantly, the proposed method of weighting sample pairs adaptively (*i.e.* WEP loss) improves the overall performance effectively.

4.8. Visualization

We show top-5 retrieved images from the FI dataset. As shown in Fig. 6(b), the results are obtained by utilizing N-pair loss to embed features. For the first two query images, the retrieved results contain several negative sentiment images, which may greatly impact the user experience. The results of the proposed method are shown in Fig. 6(c). The last two query images all obtain the correct feedback in top 5 results. Nevertheless, there is one failure case in the rank-5 result for the first query image. As we can see, though the failure image belongs to the contentment category, it also brings positive effect to the viewer's emotion, which is consistent with the polarity of the query image.

We present some attention visualization results of samples in Fig. 7. The polarity-specific attention considers the distinct color or texture details which can represent certain emotional tendency. Although these regions scatter in the image, they carry significant information which contributes to the specific emotion involved in the image. In the first image, the polarity-specific attention regions cover a great mass of blood. It guides to disgust emotion as the cue and enhance the high-level attention features in some ways. The ragged and shabby wall in the second image is attended by polarity-specific attention, while the region containing the person is drawn more attention in the emotion-specific attention map. Therefore, the polarity-specific attention can supplement this deficiency of emotion-specific attention.

5. Conclusion

In this paper, we propose an attention-aware polarity sensitive embedding network for affective image retrieval. The polarity- and emotion-specific attended features are integrated effectively. We present a weighted emotion-pair (WEP) loss, which constrains features from inter- and intra-polarity respectively. Then the sample pairs are weighted based on confidence scores derived from attention module adaptively. Finally, the total loss consisting of WEP and attention losses is exploited to optimize the architecture. Extensive experiments on four datasets indicate that our method outperforms the state-of-the-art approaches.

Acknowledgment

This work was supported by the NSFC (No.61876094, 61701273, U1933114), Natural Science Foundation of Tianjin, China (No.18JCYBJC15400, 18ZXZNGX00110), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR), the Fundamental Research Funds for the Central Universities, and Berkeley DeepDrive.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[2] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*, 2013.

[3] Manuel G. Calvo and Peter J. Lang. Gaze patterns when looking at emotional pictures: Motivationally biased attention. *Motivation and Emotion*, 28(3):221–243, 2004.

[4] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, 2016.

[5] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 2016.

[6] Ming Chen, Lu Zhang, and Jan P. Allebach. Learning deep features for image emotion classification. In *ICIP*, 2015.

[7] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. DeepSentiBank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*, 2014.

[8] Tianlang Chen, Zhongping Zhang, Quanzeng You, Chen Fang, Zhaowen Wang, Hailin Jin, and Jiebo Luo. Factual or emotional: Stylized image captioning with adaptive learning and attention. In *ECCV*, 2018.

[9] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. In *CVPR*, 2017.

[10] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005.

[11] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. *arXiv preprint arXiv:1702.07432*, 2017.

[12] Rebecca J Compton. The interface between emotion and attention: A review of evidence from psychology and neuroscience. *Behavioral and Cognitive Neuroscience Reviews*, 2(2):115–129, 2003.

[13] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[15] Yueqi Duan, Wenzhao Zheng, Xudong Lin, Jiwen Lu, and Jie Zhou. Deep adversarial metric learning. In *CVPR*, 2018.

[16] Yue Gao, Meng Wang, Dacheng Tao, Rongrong Ji, and Qionghai Dai. 3-D object retrieval and recognition with hypergraph analysis. *IEEE Transactions on Image Processing*, 21(9):4290–4303, 2012.

[17] Weifeng Ge, Weilin Huang, Dengke Dong, and Matthew R Scott. Deep metric learning with hierarchical triplet loss. In *ECCV*, 2018.

[18] Ben Harwood, Vijay Kumar BG, Gustavo Carneiro, Ian Reid, and Tom Drummond. Smart mining for deep metric learning. In *ICCV*, 2017.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[20] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Sharable and individual multi-view metric learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(9):2281–2288, 2018.

[21] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[22] Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *ACM MM*, 2015.

[23] Hye-Rin Kim, Yeong-Seok Kim, Seon Joo Kim, and In-Kwon Lee. Building emotional machines: recognizing image emotions through deep neural networks. *IEEE Transactions on Multimedia*, 20(11):2980–2992, 2018.

[24] Youngrae Kim, Yunhee Shin, So-jung Kim, Eun Yi Kim, and Hyoseop Shin. EBIR: Emotion-based image retrieval. In *ICCE*, 2009.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[26] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. In *CVPR*, 2015.

[27] David G Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[28] Jiwen Lu, Junlin Hu, and Yap-Peng Tan. Discriminative deep metric learning for face and kinship verification. *IEEE Transactions on Image Processing*, 26(9):4269–4282, 2017.

[29] Xin Lu, Poonam Suryanarayan, Reginald B. Adams Jr, Jia Li, Michelle G. Newman, and James Z. Wang. On shape and the computability of emotions. In *ACM MM*, 2012.

[30] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM MM*, 2010.

[31] Alexander Patrick Mathews, Lexing Xie, and Xuming He. SentiCap: Generating image descriptions with sentiments. In *AAAI*, 2016.

[32] Joseph A. Mikels, Barbara L. Fredrickson, Gregory R. Larkin, Casey M. Lindberg, Sam J. Maglio, and Patricia A. Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior Research Methods*, 37(4):626–630, 2005.

[33] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, 2016.

[34] Katarzyna Agnieszka Olkiewicz and Urszula Markowska-Kaczmar. Emotion-based image retrievalan artificial neural network approach. In *IMCSIT*, 2010.

[35] Rameswar Panda, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K Roy-Chowdhury. Con-

templating visual emotions: Understanding and overcoming dataset bias. In *ECCV*, 2018.

[36] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135, 2008.

[37] Lei Pang, Shiai Zhu, and Chong-Wah Ngo. Deep multimodal learning for affective analysis and retrieval. *IEEE Transactions on Multimedia*, 17(11):2008–2020, 2015.

[38] Christopher J. Patrick, Bruce N. Cuthbert, and Peter J. Lang. Emotion in the criminal psychopath: fear image processing. *Journal of Abnormal Psychology*, 103(3):523, 1994.

[39] Shengsheng Qian, Tianzhu Zhang, and Changsheng Xu. Multi-modal multi-view topic-opinion mining for social event analysis. In *ACM MM*, 2016.

[40] Tianrong Rao, Xiaoxu Li, Haimin Zhang, and Min Xu. Multi-level region-based convolutional neural network for image emotion classification. *Neurocomputing*, 333:429–439, 2019.

[41] Tianrong Rao, Min Xu, and Dong Xu. Learning multi-level deep representations for image emotion classification. *arXiv preprint arXiv:1611.07145*, 2016.

[42] Andreza Sartori, Dubravko Culibrk, Yan Yan, and Nicu Sebe. Who's afraid of Itten: Using the art theory of color combination to analyze emotions in abstract paintings. In *ACM MM*, 2015.

[43] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[44] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[46] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 2016.

[47] Kaikai Song, Ting Yao, Qiang Ling, and Tao Mei. Boosting image sentiment analysis with visual attention. *Neurocomputing*, 312:218–228, 2018.

[48] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[49] Patricia Valdez and Albert Mehrabian. Effects of color on emotions. *Journal of Experimental Psychology: General*, 123(4):394, 1994.

[50] Patrik Vuilleumier. How brains beware: neural mechanisms of emotional attention. *Trends in Cognitive Sciences*, 9(12):585–594, 2005.

[51] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, 2017.

[52] Jingwen Wang, Jianlong Fu, Yong Xu, and Tao Mei. Beyond object recognition: Visual sentiment analysis with deep coupled adjective and noun neural networks. In *IJCAI*, 2016.

[53] Xinshao Wang, Yang Hua, Elyor Kodirov, Guosheng Hu, and Neil M Robertson. Deep metric learning by online soft mining and class-aware attention. *arXiv preprint arXiv:1811.01459*, 2018.

[54] Wang Wei-ning, Yu Ying-lin, and Jiang Sheng-ming. Image retrieval by emotional semantics: A study of emotional space and feature extraction. In *SMC*, 2006.

[55] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *CVPR*, 2018.

[56] Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L. Rosin, and Ming-Hsuan Yang. Weakly supervised coupled networks for visual sentiment analysis. In *CVPR*, 2018.

[57] Jufeng Yang, Dongyu She, Yu-Kun Lai, and Ming-Hsuan Yang. Retrieving and classifying affective images via deep metric learning. In *AAAI*, 2018.

[58] Jufeng Yang, Dongyu She, and Ming Sun. Joint image emotion classification and distribution learning via deep convolutional neural network. In *AAAI*, 2017.

[59] Jufeng Yang, Dongyu She, Ming Sun, Ming-Ming Cheng, Paul L. Rosin, and Liang Wang. Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Transactions on Multimedia*, 20(9):2513–2525, 2018.

[60] Victoria Yanulevskaya, Jan C. van Gemert, Katharina Roth, Ann-Katrin Herbold, Nicu Sebe, and Jan-Mark Geusebroek. Emotional valence categorization using holistic image features. In *ICIP*, 2008.

[61] Quanzeng You, Liangliang Cao, Hailin Jin, and Jiebo Luo. Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks. In *ACM MM*, 2016.

[62] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*, 2015.

[63] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *AAAI*, 2016.

[64] Yuhui Yuan, Kuiyuan Yang, and Chao Zhang. Hard-aware deeply cascaded embedding. In *ICCV*, 2017.

[65] He Zhang, Zhirong Yang, Mehmet Gönen, Markus Koskela, Jorma Laaksonen, Timo Honkela, and Erkki Oja. Affective abstract image classification and retrieval using multiple kernel learning. In *ICONIP*, 2013.

[66] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *CVPR*, 2018.

[67] Sicheng Zhao, Guiguang Ding, Yue Gao, and Jungong Han. Learning visual emotion distributions via multi-modal features fusion. In *ACM MM*, 2017.

[68] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. Exploring principles-of-art features for image emotion recognition. In *ACM MM*, 2014.

[69] Sicheng Zhao, Chuang Lin, Pengfei Xu, Sendong Zhao, Yuchen Guo, R. V. V. Murali Krishna, Guiguang Ding, and Kurt Keutzer. CycleEmotionGAN: Emotional semantic consistency preserved CycleGAN for adapting image emotions. In *AAAI*, 2019.

[70] Sicheng Zhao, Hongxun Yao, Yue Gao, Guiguang Ding, and Tat-Seng Chua. Predicting personalized image emotion perceptions in social networks. *IEEE Transactions on Affective Computing*, 9(4):526–540, 2018.

[71] Sicheng Zhao, Hongxun Yao, Xiaolei Jiang, and Xiaoshuai Sun. Predicting discrete probability distribution of image emotions. In *ICIP*, 2015.

[72] Sicheng Zhao, Hongxun Yao, You Yang, and Yanhao Zhang. Affective image retrieval via multi-graph learning. In *ACM MM*, 2014.

[73] Sicheng Zhao, Xin Zhao, Guiguang Ding, and Kurt Keutzer. EmotionGAN: Unsupervised domain adaptation for learning discrete probability distributions of image emotions. In *ACM MM*, 2018.

[74] Xinge Zhu, Liang Li, Weigang Zhang, Tianrong Rao, Min Xu, Qingming Huang, and Dong Xu. Dependency exploitation: a unified CNN-RNN approach for visual emotion recognition. In *IJCAI*, 2017.