APSE: Attention-aware Polarity-Sensitive Embedding for Emotion-based Image Retrieval

Xingxu Yao, Sicheng Zhao, Yu-Kun Lai, Dongyu She, Jie Liang, Jufeng Yang

***Abstract*—With the popularity of social media, an increasing number of people are accustomed to expressing their feelings and emotions online using images and videos. An emotion-based image retrieval (EBIR) system is useful for obtaining visual contents with desired emotions from a massive repository. Existing EBIR methods mainly focus on modeling the global characteristics of visual content without considering the crucial role of informative regions of interest in conveying emotions. Further, they ignore the hierarchical relationships between coarse polarities and fine categories of emotions. In this paper, we design an attention-aware polarity-sensitive embedding (APSE) network to address these issues. First, we develop a hierarchical attention mechanism to automatically discover and model the informative regions of interest. Specifically, both polarity- and emotion-specific attended representations are aggregated for discriminative feature embedding. Second, we propose a generated emotion-pair (GEP) loss to simultaneously consider the inter- and intra-polarity relationships of the emotion labels. Moreover, we adaptively generate negative examples of different hard levels in the feature space guided by the attention module to further improve the performance of feature embedding. Extensive experiments on four popular benchmark datasets demonstrate that the proposed APSE method outperforms the state-of-the-art EBIR approaches by a large margin.**

I. Introduction

Images can vividly convey rich opinions and feelings of people, especially those posted on social media such as Instagram[1] and Flickr[2]. In the past few years, visual emotion analysis has attracted increasing attention in the fields of both psychology [2], [3] and multimedia [4], [5]. The related research findings can be applied in various domains, including opinion mining [6], [7], [8], psychological health [9], [10], business intelligence [11], [12], entertainment [13], [14], *etc*.

Emotion-based image retrieval (EBIR) aims to retrieve images that evoke similar emotions to the query image. Compared with content-based image retrieval (CBIR), EBIR is mainly concerned with the domain of abstract emotional semantics and subjective human perceptions, in which a so-called *affective gap* [15] exists between low-level image features and high-level abstract emotions. For this significant yet challenging task, previous studies [16], [17], [18] have made great efforts to design robust EBIR systems. To bridge the gap, in earlier years, various hand-crafted visual features were developed, inspired by the theories of psychology and art [19], [20]. In [21], Zhao *et al*. utilized multi-graph learning for EBIR based on the features of different levels, including color, attributes, facial expressions, *etc*. Recently, with the rapid development of deep learning, convolutional neutral network (CNN)-based methods have emerged that map emotional features into measurable space [22], [23]. Yang *et al*. [24] designed a joint CNN-based framework to simultaneously optimize the emotion classification and retrieval tasks, leading to performance improvements on both tasks.

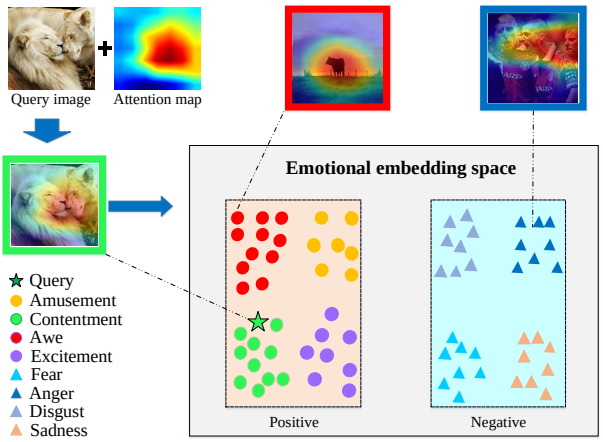


Fig. 1. Illustration of retrieving affective images in the embedding space. The two regions in the space represent binary sentiment polarities, *i.e*., positive and negative. For the given query image in a green box, the images from the same polarity but different category and from the opposite polarity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

However, two essential characteristics of image emotion are ignored in the existing EBIR methods. First, some attractive regions of an image play a decisive role in evoking emotions [25]. As shown in Fig. 1, the emotions of different samples are largely determined by the attended content of the heat maps. For example, the two face-to-face lions in the query image convey the contentment emotion due to the close proximity of their faces. Second, there exist obvious hierarchical relations among different emotions, as depicted by the embedding space in Fig. 1. We can simply classify the emotion of images based on the polarity, *i.e*., *positive* and *negative*, in the coarse level. Furthermore, as defined in psychological

X. Yao and J. Yang are with the College of Computer Science, Nankai University, China (e-mail: yxx_hbgd@163.com; yangjufeng@nankai.edu.cn).
S. Zhao is with the Department of Electrical Engineering and Computer Sciences, University of California Berkeley, U.S.A (e-mail: schzhao@gmail.com).
Y.-K. Lai is with the School of Computer Science and Informatics, Cardiff University, U.K (e-mail: LaiY4@cardiff.ac.uk).
D. She is with the Department of Computer Science, Tsinghua University, China (e-mail: sherry6656@163.com).
J. Liang is with the Department of Computing, The Hong Kong Polytechnic University, HKSAR (e-mail: liang27jie@163.com).
A preliminary version of this work appeared at ICCV [1].
[1]https://www.instagram.com/
[2]https://www.flickr.com/

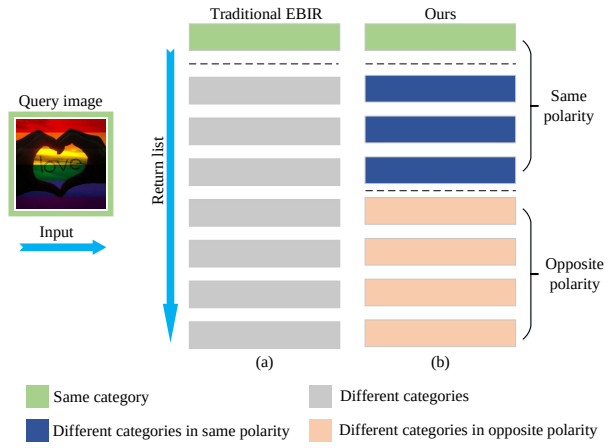


Fig. 2. Illustration of the expected rank list. (a) Ranking list without considering the hierarchy of emotions as done in traditional EBIR methods. (b) Ranking list of optimization objective in this paper.

theories [3], [26], we are also able to recognize emotions at a more concrete level, *i.e*., *amusement, contentment, awe, excitement, fear, anger, disgust, and sadness*. The first four categories belong to the 'positive' polarity, while the last four categories belong to 'negative' polarity. In this paper, we use the term 'class' to represent both polarity and emotion categories, where 'category' means the concrete emotions. When measuring the emotional similarity, we need to consider not only the emotion category but also polarity because categories belonging to the same polarity are more similar than those belonging to the opposite polarity. Explicitly, our objective (as shown in Fig. 2) is to rank the images in a gallery based on the relationship with the query image in the following order: the same emotional category, the same polarity but different emotion categories, and different polarities.

To consider the emotional characteristics mentioned above, in this paper, we propose an attention-aware polarity-sensitive embedding (APSE) network for EBIR. An attention module is used to attend to emotion-related regions. While concrete emotion categories depend on high-level semantic information, polarity is relevant to low-level features such as color, texture, *etc*. [27], [28], [20]. Consequently, in the attention module, we utilize *polarity-specific* attention in lower layers, while *emotion-specific* attention is utilized in higher layers. Then, the two types of attended features are integrated by cross-level bilinear (CLB) pooling, which can facilitate the interaction between the information of different levels. The polarity sensitivity is not only reflected in our attention module but also taken into account in the embedding learning. In particular, we propose to optimize a new generated emotion-pair (GEP) loss, which is designed based on the N-pair loss [29], to learn discriminative feature embedding. First, the samples in the embedding space are separated into two parts based on their polarities (negative and positive). This is mainly because the primary goal of EBIR is to successfully retrieve the images with the same polarity as the query. Second, the different categories in the same polarity can also be well distinguished in the objective function. In addition, the hardness of negative examples is augmented by generating embedding with different degrees based on the category probability in the attention module. With the generated hard negative feature embedding, not only can the model convergence be accelerated but also, more importantly, the performance of embedding learning is improved. During the end-to-end training process, the unified framework simultaneously optimizes the GEP loss and attention loss to map raw images into emotional feature embeddings used for EBIR.

Our contributions are highlighted as follows:

- We propose multi-level attended local features for EBIR based on psychology theories that indicate low-level and high-level image features are relevant to different levels of the emotion hierarchy. To the best of our knowledge, we are the first to integrate attended features at different levels to capture emotional information.
- We develop an attention-aware polarity sensitive embedding (APSE) network that takes into account the inter- and intra-polarity relationships of the emotion labels. The proposed GEP loss connects the attention module and feature embedding effectively during the training process. Extensive experiments indicate that the proposed architecture significantly outperforms the state-of-the-art methods on four benchmark datasets.

This journal paper improves on our preliminary conference version [1] in the following three aspects. (1) We develop a method that adaptively generates harder negative examples in the embedding learning process that can learn more discriminative features. (2) We provide more implementation details and sufficient visualization results to showcase the effectiveness of the proposed method and provide more insights regarding the key essence of an EBIR system. Moreover, we systematically discuss the failure cases and show more experimental results, including experiments that elaborate on the choice of feature combinations at different levels. (3) A more comprehensive survey of related work is performed, and the performance of the latest methods is supplemented in comparison experiments.

The rest of the paper is organized as follows. Section II summarizes the related work on image emotion analysis, visual attention mechanisms, and deep feature embedding. Section III introduces the proposed hierarchical attention mechanism and polarity-sensitive embedding learning method. In Section IV, we perform both quantitative and qualitative experiments on popular benchmark datasets and analyze the results. Finally, Section V concludes this paper.

II. Related Work

In this section, we review closely related work in the past decades, including image emotion analysis [30], [31], [32], the visual attention mechanism [33], [34], and feature embedding learning [35], [36].

A. Image Emotion Analysis

In the domain of image emotion analysis, most of the studies focus on dominant emotion classification [37], [38], [39] and emotion distribution learning [24]. In the early years,

various hand-crafted features were introduced, inspired by the theories of art and psychology [19], [20]. The effectiveness of low-level features [19], [28], such as *color, texture, shape*, *etc*., and mid-level representations [40], such as *attribute, principle-of-the-art features*, *etc*., were demonstrated when representing emotion at that time. To better bridge the "affective gap" between low-level representations and abstract emotion semantics, Borth *et al*. [41] proposed adjective-noun pairs (ANP) such as "beautiful flower" to describe an image. In addition, facial expressions [42] act as a very important element for recognizing emotions, as demonstrated in [19]. Along with the boom of deep learning methods, an increasing number of researchers [23], [43] have utilized images to train CNNs for specific image emotion analysis tasks. With the supervision of the emotion labels, the learned features can well capture the characteristic representation for each category [44]. Moreover, considering that producing emotion is relevant to various visual stimuli from a low level to a high level, some studies [30], [45] extracted features from multiple layers to obtain more comprehensive information. Further, Zhu *et al*. [46], [47] explored the dependency between features of different levels by employing the bidirectional gated recurrent unit (Bi-GRU).

Although EBIR is meaningful to the affective computing community and has many applications, it has drawn less attention than emotion recognition. As presented in the discussion in [48], EBIR can be regarded as a branch of semantic-based image retrieval as well as emotion recognition. EBIR was first proposed in [17] in which a system was constructed to support the query interface of two types, including emotional keywords and emotional images. To capture reasonable emotion characteristics, the authors extracted color, texture, and pattern as clues. Obviously, this method highly depends on the quality of hand-designed emotion features. Afterward, Olkiewicz *et al*. [49] used an artificial neural network to extract emotional features for EBIR. Further, the authors also exploited the retrieval results to label images with emotional keywords. Beyond a single modality, Xing *et al*. [50] explored an interesting multi-modality task of emotion-driven Chinese folk music-image retrieval, developing a new perspective for EBIR. However, in this paper, we only focus on emotional image retrieval. In the last few years, Zhao *et al*. [21] retrieved emotional images by employing multi-graph learning, where each graph contains one type of hand-crafted feature. Inspired by the deep Bolzmann machine (DBM), Pang *et al*. [51] developed a density model to learn the joint representation coupled with emotions and semantics, which can be used for emotion-oriented cross-modal retrieval. With the emergence of CNNs, a unified multi-task framework [24] has been designed to simultaneously learn retrieval and classification tasks. However, existing EBIR methods fail to fully employ important cues such as information from multiple levels or the hierarchy of emotional labels. In this paper, we develop a polarity-sensitive embedding method based on multi-level attended features for EBIR.

B. Visual Attention Mechanism

Imitating human attention, we expect that a network can weight features by their degree of importance for a task and thus further obtain more discriminative features. The effectiveness of the attention mechanism has been demonstrated in various visual tasks, including image captioning [52], [53], person re-identification [54], object detection [55], *etc*. In [56], a residual attention network was proposed by incorporating soft attention into the state-of-the-art CNN architecture. Self-attention [57] has been used to compute the response of one position through attending all positions. In computer vision, attention is able to capture the dependency of different regions in the same image. As the extension of self-attention in visual tasks, non-local networks [58] can capture the long-range dependency by calculating the interactions between two frames of a video or two regions of an image.

Based on theories of psychology [59], [60], emotional contents, including smiling faces, cute babies, beautiful flowers, *etc*., always elicit more attention from humans. Unlike traditional object classification and detection tasks in which object regions are explicit and well-defined, emotions are ambiguous and may contain foreground and background [61], [62]. In the early years, prior methods [30], [63] detected emotional attention regions from a large number of candidate bounding boxes by computing both an objectiveness score and an emotion score. It is obvious that these methods consume excessive amounts of time and computing resources. In [62], Fan *et al*. performed human fixation based on expensive eye-tracking data and then evaluated the relationship between image sentiment and visual stimuli. Yang *et al*. [61] proposed to directly generate soft attention maps with the single shot by weighting feature responses on various emotion categories. Differently, in this paper, we take into account the representations from multiple layers and develop a hierarchical attention mechanism for learning discriminative features 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.

C. Feature Embedding Learning

In the past years, various metric learning methods [64], [36] have been proposed to learn feature embedding in a separate space, and they have a wide range of applications in the domain of computer vision [65], [66]. The most representative metric learning loss functions are contrastive loss [67] and triplet loss [68], which later motivated a variety of novel methods. The contrastive loss aims to minimize the distance between samples of the same class and push away the samples of different classes with a fixed margin. The triplet samples include the anchor, positive, and negative examples. The triplet loss encourages that the distance between the anchor and the negative example is larger than that between the anchor and the positive example by at least a specified margin. As an extension of the contrastive loss, a lifted embedding structure [69] is proposed to compute the loss based on the matrix consisting of pairwise distances of the mini-batch. Beyond the triplets, Chen *et al*. [65] introduced the negative

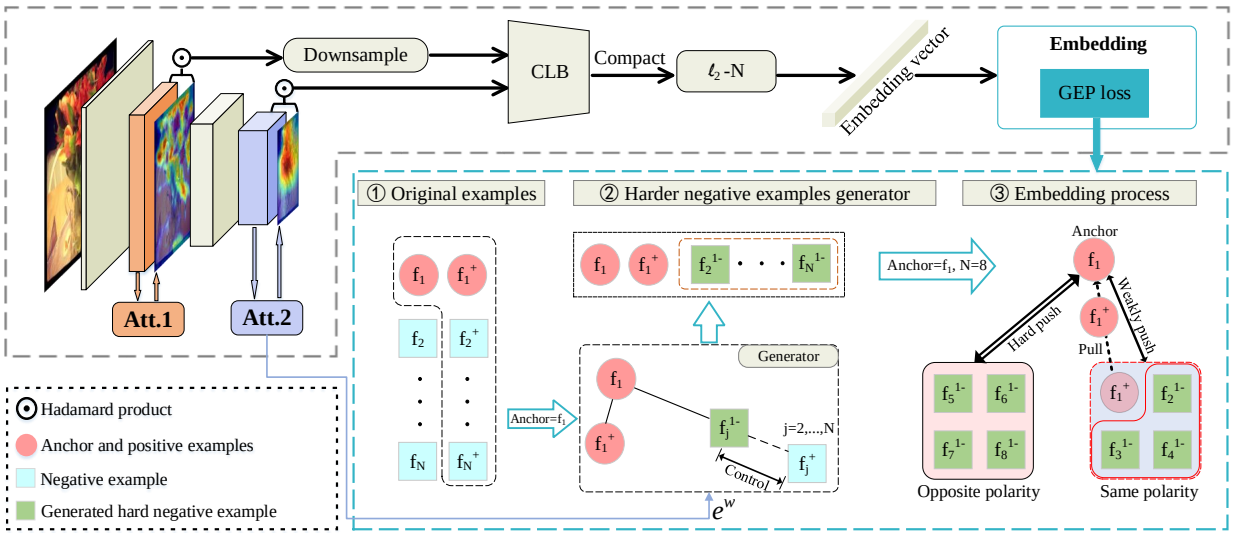


Fig. 3. Pipeline of the proposed approach. The attended features that are output from attention modules (Att.1 and Att.2) of different levels are integrated by CLB. After compaction and ℓ_2-Normalization (ℓ_2-N), the combined representations are input into embedding space for metric learning. In the GEP loss, we employ similar emotion categories in the FI dataset [26] with the number of categories $N = 8$. Here, four categories are positive and the other four are negative. The detailed process of generating attention maps is presented in Fig. 4. Att.1 and Att.2 represent polarity-specific attention and emotion-specific attention, respectively. f_i and f_i^+ represent the features of the anchor point and positive example from the i^{th} category, respectively. f_j^{i-} means the generated negative embedding of the j^{th} category for the anchor from the i^{th} category.

pairs w.r.t. different probe samples. Furthermore, to generalize the application of metric learning to continuous labels, Kim *et al*. [64] proposed a log-ratio loss to learn feature embedding based on the distance between labels. The method can be well applied to the task in which the labels are continuous, such as human pose estimation, considering a novel relation of samples.

In metric learning, the sampling strategy may affect the training process. Therefore, some studies aim to design effective sampling strategies to accelerate the convergence and obtain better performance. To select the informative triplets that violate the constraints, an online negative sample mining strategy was proposed in [68], including the hardest negative mining and semi-hard negative mining. Moreover, Duan *et al*. [70] generated hard negatives by deep adversarial learning to train a more discriminative model. Considering that the prior mining methods cannot well characterize the global geometry of the embedding space, hardness-aware metric learning [71] has been proposed to adaptively generate samples with different hard levels based on the training status.

The concept of hierarchy has been discussed in some retrieval tasks. To effectively retrieve fashion products, Liao *et al*. [72] constructed an EI (exclusive& independent) tree that models hierarchical structures based on product taxonomy and domain knowledge. Based on the EI tree, they further proposed a hierarchical similarity function based on triplet loss to characterize the semantic similarities among fashion products. For the complicated relation between different products, it is difficult to adjust the optimal margin for triplets. The suboptimal margin may result in slow convergence and suboptimal results. Similarly, considering the hierarchical relation of labels, Wang *et al*. [73] first proposed supervised hierarchical deep hashing by weighting each layer in the tree structure that describes the semantic of labels. In [74], Peng *et al*. considered the hierarchy that exists in instances for cross-modal retrieval. Specifically, the authors fused the embedding of coarse-grained instances and fine-grained patches using two pathway networks to make cross-modal correlation more precise. In this work, we focus on the hierarchy of the emotional label.

Motivated by the observation that there is an obvious hierarchy in emotion labels, *i.e*., from coarse polarity to concrete emotions, we design a polarity-sensitive GEP loss for optimizing our framework. The most similar work to ours is [35], which constructs a hierarchical structure based on the triplet loss. Unlike this method that needs to use a special sampling strategy, our method can directly take full advantage of all the samples within a mini-batch, avoiding the redundant computations.

III. Methodology

We design a novel network, named the attention-aware positive embedding (ASPE) network, to learn feature embedding for emotional images. The framework contains two main closely related components, as shown in Fig. 3. One is the hierarchical attention module that integrates polarity- and emotion-specific attended features extracted from multiple layers (Sec. III-A); the other is the embedding module that learns polarity-sensitive feature embedding by optimizing GEP loss guided by the attention module (Sec. III-B).

A. Hierarchical Attention Mechanism

We introduce a simple yet effective attention module that detects informative regions for different hierarchies of emotion labels in both higher and lower layers. As shown in Fig. 4,

there are two components in our attention module, *i.e.*, attention head and output head. In the attention head, the extracted feature maps are first weighted by spatial attention and then are reduced to K dimensions. With the supervision of attention loss, the score of each feature map can learn the feature activation of the corresponding class (polarity or emotion). In the output head, the final attention map is generated by computing the sum of the feature maps of all classes weighted by corresponding score results from the attention head. Note that the attention module can be applied in multiple layers.

Suppose that we conduct attention in the l^{th} layer for instance. Its feature maps $F^l \in \mathbb{R}^{h\times w\times c}$ from the l^{th} convolutional layer will be fed into the attention head, and then K^l attention maps derived from F^l are output. h, w and c represent the height and width of the feature maps, and the number of channels, respectively, while K^l denotes the number of labels in the l^{th} layer. In the lower layers that are supervised by binary sentiment polarities, the value of K is set to 2, while we set $K = 8$ in higher layers, representing eight specific emotion categories as defined in Mikel's wheel [3]. In the spatial attention, we intend to consider the emotion-related regions rather than treating each region equally. Thus, we aggregate the received feature activation tensor in a channel-wise approach and then feed the derived 2-D aggregated maps into a softmax layer. We formulate the process as:

$$Z^l = \text{Softmax}(\sum_{i=1}^{c} F_i^l), \tag{1}$$

where Z^l is the output of spatial weights, and F_i^l is the feature map of the i^{th} channel.

Then, we conduct spatial attention on feature maps to compute the spatially attended feature maps, *i.e.*, $\hat{F}^l = F^l \odot Z^l$, where $\odot$ means the Hadamard Product by repeating Z^l for each channel of F^l. After generating $\hat{F}^l$, a 1×1 conv. layer is employed to reduce the channel-wise dimension from c to K^l, resulting in $S^l \in \mathbb{R}^{h\times w\times K^l}$. In S^l, each 2-dimensional feature map represents a sentiment polarity or specific emotion category, which depends on the value of l. Then, S^l is fed into a global average pooling (GAP) layer and a softmax layer successively, acquiring a confidence score $\boldsymbol{c}^l$, in which each element that represents global information for each feature map ranges from 0 to 1 and the sum of them is 1.

In the output head of the l^{th} level, the 2-dimensional class-wise feature maps S^l and the derived confidence score vector $\boldsymbol{c}^l$ for different classes are input. Note that each element c_j^l in the confidence vector usually well represents the degree that the j^{th} label describes the instance. To comprehensively consider the responses for different classes when computing the final attention weights, we add all the class-wise feature maps weighted by their corresponding scores. Therefore, the specific process of generating attention map $\mathcal{U}$ can be formalized as follows (with the layer-wise subscript l omitted without ambiguity):

$$\mathcal{U} = norm(\sum_{j=1}^{K} c_j S_j), \tag{2}$$

where $norm$ represents the normalization on the 2-dimensional attention map, and S_j denotes the feature activations for the j^{th} label. Then, to obtain the final attended features F^a, we apply the attention weights $\mathcal{U}$ on the feature maps $\hat{F}$ derived from the attention head: $F^a = \hat{F} \odot \mathcal{U}$, where $\odot$ denotes element-wise multiplication by broadcasting. In practice, we train our network by imposing constraints using the labels of different hierarchies in different layers. Therefore, the attention loss in different layers can be represented in the following unified formula:

$$\mathcal{L}_{att} = -\frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{K} \mathbf{1}[z_m = j] \log c_j, \tag{3}$$

where $\mathbf{1}[t] = 1$ if the condition t is true, and 0 otherwise. M represents the total number of input images, and z_m is the corresponding label ID for the m^{th} input image. Particularly, we simultaneously employ the loss function on both lower and higher layers, resulting in attention weights for two polarities and eight emotion categories.

Since the attended features from different layers focus on different aspects [45], [46], we intend to effectively integrate these various sources of information for a more discriminative representation. To multiply attended features of different scales through CLB, we first downsample $F^{a1} \in \mathbb{R}^{h^1\times w^1\times k^1}$ output from lower layers to $\overline{F^{a1}} \in \mathbb{R}^{h^2\times w^2\times k^1}$ whose size is the same as $F^{a2} \in \mathbb{R}^{h^2\times w^2\times k^2}$. Then, we utilize the CLB operation to model the interactions of different level features and establish pairwise correlations between the channels. Let $L' = h^2 \cdot w^2$ be the number of locations in the feature map. In the u-th location, we evaluate the channel-wise matrix outer product between $\overline{F_u^{a1}} \in \mathbb{R}^{1\times k_1}$ and $F_u^{a2} \in \mathbb{R}^{1\times k_2}$. Then, the bilinear output representation φ can be calculated through $\sum_{u=1}^{L'} \overline{F_u^{a1}}^{\top} F_u^{a2}$. φ is reshaped from $k_1 \times k_2$ to $1 \times k_1 k_2$, followed by dimensionality reduction for a more compact feature representation. Following [75], the compressed φ is then passed successively through signed square root and ℓ_2 normalization.

B. Polarity-Sensitive Embedding Learning

In this section, to take into account the hierarchy in the emotion label space, we first introduce the polarity-sensitive emotion-pair (EP) loss based on the N-pair loss. Moreover, to enhance the robustness of the trained model, we further generate negative examples for each anchor-positive pair based on their original negative examples. The generation strategy can be adjusted by the confidence scores from the attention module.

1) Review on N-pair loss: The N-pair loss [29] is proposed based on the $(N+1)$-tuplet $\left\{x, x^+, x_1^-, \cdots, x_{N-1}^-\right\}$, including an anchor x, a positive example x^+, and $N-1$ negative examples. Its aim is to identify a positive example for an anchor from all the negative examples. To fully exploit training data, N pairs of convolution features constructed from N different categories are formulated as $\left\{(f_1, f_1^+), \cdots, (f_N, f_N^+)\right\}$. Note that f_i and f_i^+ represent the feature embeddings of anchor point x_i and positive example x_i^+, respectively, both from the i^{th} category. In the feature space, f_i^+ serves as a negative example for the anchor from the j^{th} category, where $i \neq j$.

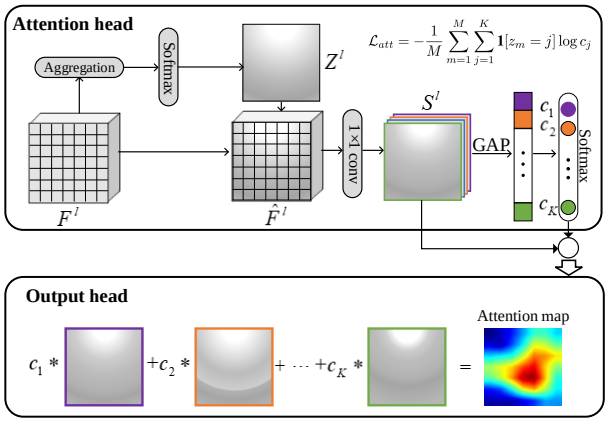


Fig. 4. Overview of our attention map generation. In the input head, F^l denotes the feature map extracted from l^{th} layer, and Z^l represents the spatial weights. After dimension reduction using 1×1 convolution, the class-aware activation S^l and corresponding confidence score c_i on the i^{th} category are derived. Note that GAP denotes global average pooling. In the output head, the resulting attention map is obtained by weighting activation maps using the corresponding confidence scores. In the lower layers, the attention module generates a polarity-specific attention map, whereas an emotion-specific attention map is generated in higher layers.

Discarding the subscript of f for simplicity, the similarity between f and f^+ has a positive correlation with the value of their dot product $f^\top f^+$. Therefore, the formula of N-pair loss is given as:

$$\mathcal{L}_{np} = \frac{1}{N}\sum_{i=1}^{N}\log(1+\sum_{j\neq i}\exp(f_i^\top f_j^+ - f_i^\top f_i^+)). \quad (4)$$

With this penalty strategy, $N-1$ negative examples are simultaneously pushed away from the anchor.

2) EP loss: Although the N-pair loss has demonstrated its effectiveness in various tasks, it is insufficient to learn the feature embeddings for emotional images well due to the negligence of sentiment polarity, *i.e.*, positive and negative. Intuitively, examples from the same polarity as the query should be closer to it than those from the opposite polarity in the embedding space. To achieve this goal, we propose an inter-polarity loss to effectively separate the two polarities. Specifically, in an N-tuple, we regard the examples from the opposite polarity as a group and compute their mean similarity to enlarge the distance from negative examples of the same polarity. Here, negative examples mean images with categories that differ from the anchor. Note that the positive examples (images from the same category as the anchor) will not contribute to the optimization of this loss function. It is mainly because a positive example can dramatically reduce the mean value of the distance between the anchor and examples of the same polarity, resulting in insufficient training on negative examples of the same polarity. Therefore, we formalize the inter-polarity loss as follows:

$$\begin{aligned}\mathcal{L}_{inter} = \frac{1}{N}\sum_{i=1}^{N}\log(1+\exp(\frac{1}{N_{\mathcal{Q}_i}}\sum_{j\in\mathcal{Q}_i} f_i^\top f_j^+ \\ - \frac{1}{N_{\mathcal{P}_i}}\sum_{j\in\mathcal{P}_i, j\neq i} f_i^\top f_j^+)),\end{aligned} \quad (5)$$

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

With the inter-polarity loss, we are able to largely avoid the dramatic failure cases that occur when many examples of the opposite polarity rank at the top of returned list, which may lead to an unpleasant experience for users. Further, it is more challenging to distinguish the positive examples from the negative examples of the same polarity. Therefore, to learn more discriminative feature embeddings, we develop an intra-polarity loss that can differentiate similar categories within the same polarity as follows:

$$\mathcal{L}_{intra} = \frac{1}{N}\sum_{i=1}^{N}\log(1+\sum_{j\in\mathcal{P}_i, j\neq i}\exp(f_i^\top f_j^+ - f_i^\top f_i^+)). \quad (6)$$

Then, we combine inter-polarity loss and intra-polarity loss, resulting in the EP loss:

$$\mathcal{L}_{ep} = \mathcal{L}_{inter} + \mathcal{L}_{intra}. \quad (7)$$

With the combined loss function, we can realize our aim of modulating the position of feature embeddings in the separable space according to the hierarchical emotional similarity.

3) Generating negative feature embeddings: In the learning process, many tuples will be constructed for training. In fact, a majority of them may fail to contribute to the update of parameters because they lack sufficient information and produce gradients that approach 0. Furthermore, images in the same category always have a large diversity in emotion intensity, so the uniform penalty strategy may be insufficient to optimize negative examples of various hard levels. Therefore, inspired by [71], we propose to manipulate the hard level of the training tuples adaptively by generating new negative examples based on the learning status.

Given the embedding f_i of an anchor, f_i^+ and f_j^+ $(i \neq j)$ are used to represent the feature embeddings of the corresponding positive and negative examples, respectively. Based on the existing negative example f_j^+ for f_i, we can utilize linear interpolation to adjust the hardness of training data:

$$f_j^{i-} = f_i + \lambda_0(f_j^+ - f_i), \lambda_0 \in [0,1], \quad (8)$$

where f_j^{i-} denotes the generated embedding from the j^{th} category for the anchor from the i^{th} category. However, to avoid generated examples that are too close to the anchor and lead to noisy data, the minimum value for λ_0 should be larger than $\frac{d(f_i,f_i^+)}{d(f_i,f_j^+)}$. Therefore, the range of λ_0 is $(\frac{d(f_i,f_i^+)}{d(f_i,f_j^+)}, 1]$, where $d(f_i, f_i^+)$ means the distance between the anchor and positive example ($\|f_i - f_i^+\|_2$) and $d(f_i, f_j^+)$ means the distance between the anchor and negative example ($\|f_i - f_j^+\|_2$). To achieve this, a variable $\beta \in (0,1]$ is introduced as a factor to control the value of λ_0. With β ranging from 0 to 1, λ_0 ranges from $\frac{d(f_i,f_i^+)}{d(f_i,f_j^+)}$ to 1. λ_0 can be represented as the following formula:

$$\lambda_0 = \begin{cases} \beta + (1-\beta)\frac{d(f_i, f_i^+)}{d(f_i, f_j^+)}, & \text{if } d(f_i, f_j^+) > d(f_i, f_i^+) \\ 1, & \text{if } d(f_i, f_j^+) \leq d(f_i, f_i^+). \end{cases} \tag{9}$$

At the condition of $d(f_i, f_j^+) > d(f_i, f_i^+)$, the generated negative example can be expressed as:

$$\tilde{f}_{ij}^- = f_i + [\beta d(f_i, f_j^+) + (1-\beta) d(f_i, f_i^+)] \frac{f_j^+ - f_i}{d(f_i, f_j^+)}. \tag{10}$$

To assign a proper value to β adaptively, we consider the hard level of separating the corresponding anchor-negative pair. Given an anchor x_i from the i^{th} category and one of its negative samples x_j^- from the j^{th} category, we use $c_j^{x_i}$ to represent the confidence score of x_i of the i^{th} category w.r.t. the j^{th} emotional category, while $c_i^{x_j^-}$ denotes the confidence score of x_j^- of the j^{th} category w.r.t. the j^{th} emotional category. In the attention module, a higher confidence $c_j^{x_i}$ or $c_i^{x_j^-}$ denotes that the pair is harder to separate. Consequently, we aim to assign a stronger penalty term to this pair in the embedding learning by generating negative feature embeddings that are closer to the anchor. We introduce a variable w_{ij}, which indicates the difficulty to distinguish the anchor x_i and its negative example x_j^-. It is formalized as:

$$w_{ij} = \exp(c_j^{x_i}) \cdot \exp(c_i^{x_j^-}). \tag{11}$$

The larger the weight w_{ij} is, the harder it is to separate the x_i and x_j^-, so we should impose a stronger penalty on them by generating examples that are closer to the positive examples in the feature space. To achieve this goal, we intuitively set β to $e^{-w_{ij}}$. By controlling the value of β with $e^{-w_{ij}}$, for the anchor-negative pair with higher similarity, our algorithm will generate the negative example that is closer to anchor. Therefore, the proposed algorithm that generates the negative feature embedding can be formulated as:

$$f_j^{i-} = \begin{cases} f_i + \left[e^{-w_{ij}} d(f_i, f_j^+) + (1 - e^{-w_{ij}}) d(f_i, f_i^+)\right] & \\ \qquad \text{if } d(f_i, f_j^+) > d(f, f_i^+) & \\ f_j^+, \qquad \text{if } d(f_i, f_j^+) \leq d(f, f_i^+). & \end{cases} \tag{12}$$

Consequently, in our EP loss function, the generated features are regarded as the negative examples, so we introduce GEP loss:

$$\begin{aligned} \mathcal{L}_{gep} = \frac{1}{N} \sum_{i=1}^{N} \log[(1 + \exp(\frac{1}{N_{\mathcal{Q}_i}} \sum_{j \in \mathcal{Q}_i} f_i^\top f_j^{i-} \\ - \frac{1}{N_{\mathcal{P}_i}} \sum_{j \in \mathcal{P}_i, j \neq i} f_i^\top f_j^{i-}))(1 + \sum_{j \in \mathcal{P}_i, j \neq i} \exp(f_i^\top f_j^{i-} \\ - f_i^\top f_i^+))]. \end{aligned} \tag{13}$$

We define the total loss consisting of the attention and GEP losses to optimize the proposed framework simultaneously:

$$\mathcal{L}_{total} = \lambda \mathcal{L}_{gep} + (1-\lambda) \mathcal{L}_{att}, \tag{14}$$

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

IV. Experiments

In this section, we present extensive experimental results on widely used benchmark datasets to evaluate the effectiveness of our algorithm. Apart from comprehensive comparison experiments against the state-of-the-art methods, we also conduct an ablation study to analyze each module. Finally, various visualization results are provided.

A. Datasets

We conduct our experiments on four benchmark datasets, including a large-scale dataset, *i.e.*, Flickr and Instagram (FI) [26], and three small-scale datasets, *i.e.*, Subset A of IAPS (IAPSa) [3], Artistic dataset (ArtPhoto) [19], and Abstract paintings (Abstract) [19].

1) Large-scale Dataset: FI is one of the largest well-annotated image emotion datasets, which is collected from social websites by querying with Mikel's eight emotions [3] as keywords. A total of 225 AMT workers are employed to label these images. Finally, 23,308 images that receive at least three agreements of five assigned workers are used as the final clean dataset.

2) Small-scale Datasets: IAPSa includes 395 images collected from the International Affective Picture System (IAPS) [91] and is labeled with eight emotion categories by 20 undergraduate participants. Abstract contains 228 peer-rated abstract paintings in which the color and texture occupy the major visual contents, lacking specific semantic information. Artphoto is composed of 806 artistic photos downloaded from an art sharing site. The emotion label of each image is determined by the owner of the image.

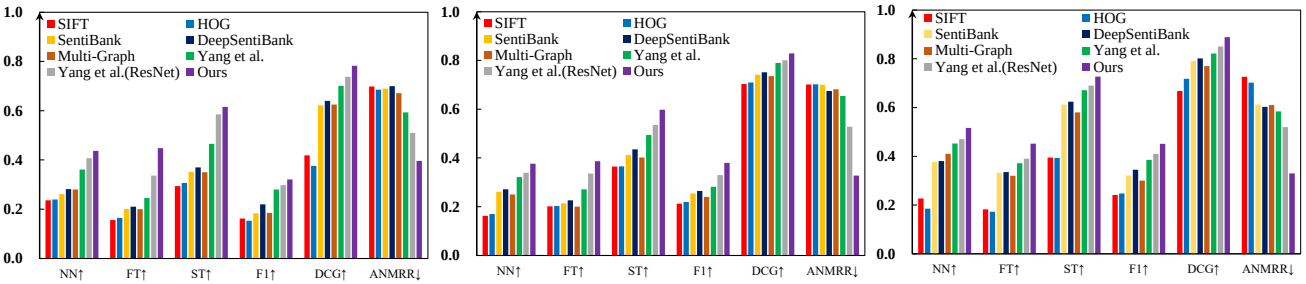
B. Evaluation Metrics

Following previous work [21], [90], we utilize the following metrics to comprehensively evaluate the experimental results. Mean average precision (mAP) is employed to measure the mean precision of retrieval results. In this paper, we consider both the mAP of eight emotion-specific categories (mAP_8) and the mAP of two sentiment polarities (mAP_2). Note that the following metrics are only used to evaluate the retrieval performance on eight specific emotions. Nearest neighbor rate (NN) represents the proportion of the rank-1 samples in the return list that are correct. First tier (FT) and second tier (ST) both denote the recall of the returned results. Specifically, FT is responsible for measuring the recall for the top-$\boldsymbol{n}$ returned results, while ST denotes the top-$2\boldsymbol{n}$ recall. Here, $\boldsymbol{n}$ is the total number of all the correct examples for the query. Assuming that users prefer frontal results, the discounted cumulative gain (DCG) [92] incorporates the weights of different positions of relevant samples in the ranking list into the performance measurement. The F_1 score is the harmonic mean of the precision and recall. Similar to DCG, the average normalized modified retrieval rank (ANMRR) [93] takes into account the ranking sequence of relevant images within the retrieved results. Note that smaller values of ANMRR represent better retrieval results, and for the other evaluation metrics, larger ones are better.

TABLE I

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' REPRESENTS THAT THE SOFTMAX LOSS IS USED FOR TRAINING, AND 'DIM.' DENOTES THE DIMENSION OF FEATURES. OLD APSE MEANS THE METHOD IN OUR CONFERENCE VERSION.

	Methods	Dim.	mAP_8 ↑	mAP_2 ↑	FT↑	ST↑	NN↑	DCG↑	ANMRR↓
TRA	SIFT [76]	1000	0.1705	0.5913	0.1830	0.3513	0.2462	0.4507	0.6553
	HOG [77]	1000	0.2115	0.6002	0.1926	0.3620	0.3225	0.4639	0.6424
	Gabor [77]	1000	0.1724	0.5942	0.1768	0.3395	0.2641	0.4434	0.6770
	SentiBank [41]	1200	0.2337	0.6168	0.2422	0.4232	0.3990	0.5223	0.5934
CNN	DeepSentiBank [78]	2089	0.2559	0.6247	0.2658	0.4468	0.4583	0.5509	0.5655
	MVSO [79]	4342	0.2798	0.6366	0.2877	0.4761	0.5158	0.5731	0.5346
	AlexNet (S) [80]	4096	0.2709	0.6328	0.2795	0.4693	0.5038	0.5633	0.5463
	VGGNet (S) [81]	4096	0.3013	0.6552	0.3007	0.4887	0.5511	0.5860	0.5161
	GoogLeNet (S) [82]	2048	0.3583	0.6773	0.3571	0.5619	0.5816	0.6403	0.4517
	ResNet (S) [83]	2048	0.4380	0.7068	0.4286	0.6079	0.6084	0.6816	0.3998
	WSCNet [61]	2048	0.5060	0.7381	0.4653	0.6223	0.6358	0.6910	0.3872
EMB	Contrastive loss [67]	2048	0.3842	0.6972	0.3768	0.5702	0.5711	0.6508	0.4396
	Triplet loss [68]	2048	0.5130	0.7120	0.4864	0.6216	0.5710	0.6843	0.3860
	N-pair loss [29]	2048	0.5217	0.8062	0.4785	0.7075	0.5341	0.7310	0.3089
	Center loss [84]	2048	0.5021	0.6943	0.4982	0.6082	0.5431	0.6789	0.3621
	Binomial deviance [85]	2048	0.5421	0.7352	0.4781	0.7112	0.5371	0.7031	0.3398
	ArcFace [86]	2048	0.5308	0.6910	0.5366	0.6675	0.6187	0.7232	0.3123
	SphereFace [87]	2048	0.4987	0.6689	0.4032	0.6023	0.6065	0.6755	0.3604
	FastAP [88]	2048	0.5639	0.7123	0.5578	0.6822	0.6112	0.7209	0.3179
	SoftTriple [89]	2048	0.5712	0.7746	0.5431	0.6921	0.6210	0.7312	0.3064
	Yang *et al*. [90]	544	0.6395	0.8081	0.5995	0.7354	0.6164	0.7866	0.2518
Ours	Old APSE [1]	512	0.7344	**0.9079**	0.6985	0.7817	0.6613	0.8114	0.2201
	New APSE	512	**0.7433**	0.9030	**0.7075**	**0.7994**	**0.6755**	**0.8250**	**0.2106**



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

Fig. 5. Retrieval performance on three small datasets (Artphoto, Abstract, and IAPSa). The results are derived using the model trained on the FI dataset.

C. Baselines

In the comparison experiments, we compare our method to different baselines. The low-level descriptors include SIFT [76], HOG [77], and Gabor [77], the dimension of which are set to 1,000. Meanwhile, we explore the performance of mid-level features, especially those designed based on ANPs, including 1200-dimensional representations of SentiBank [41], 2089-dimensional features of DeepSentiBank [78], and more recent 4342-dimensional features of MVSO (English) [79]. For CNN-based methods, we fine-tune different architectures with the supervision of softmax loss, including AlexNet, VGGNet, GoogLeNet, and ResNet-50, in which the features of the last FC layer are extracted as the representation for embedding learning. Furthermore, with the ResNet-50 model as the backbone, we also train the networks by optimizing various metric learning losses, including contrastive loss [67], triplet loss [68], center loss [84] and N-pair loss [29], *etc*. Finally, we compare with the state-of-the-art methods of EBIR, including Yang *et al*. [90], multi-graph [21], and the previous conference version of our APSE method [1].

D. Implementation Details

Following [90], we regard the test images of the FI dataset as the query images to retrieve relevant emotional images in the training set. For small-scale datasets, we use each image to retrieve the remaining images. All the images are ranked based on the emotional similarity between them and the queries. The proposed architecture is based on pretrained ResNet-50 [83]. The original images are resized to 256 $\times$ 256 and randomly cropped to 224 $\times$ 224. The framework is optimized by SGD with the weight decay of 0.0005 and a momentum of 0.9. The initialized learning rate is set as 0.001 and dropped down one-tenth for every 40 epochs. The maximal number of epochs is 100 for fine-tuning all layers with a batch size of 32, ensuring 4 images from each of the 8 emotions. We set hyper-parameter $\lambda = 0.5$ in all our experiments, which achieves the best

TABLE II
ABLATION EXPERIMENTS ON THE FI DATASET. THE BACKBONE FRAMEWORK IS RESNET-50 PRETRAINED 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 THE 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. WHEN CLB IS NOT SELECTED, THE FEATURES FROM DIFFERENT LAYERS ARE CONCATENATED DIRECTLY.

	AT	N-pair	EP	GEP	SA	HA	CLB	SO	MO	$\text{mAP}_8 \uparrow$	$\text{mAP}_2 \uparrow$	FT↑	ST↑	NN↑	DCG↑	ANMRR↓
(a)	√							√		0.4380	0.7068	0.4286	0.6079	0.6084	0.6816	0.3998
(b)		√						√		0.5217	0.8062	0.4785	0.7075	0.5341	0.7310	0.3089
(c)			√					√		0.5680	0.8558	0.5247	0.7187	0.5623	0.7602	0.2789
(d)	√	√						√		0.6225	0.7816	0.5779	0.7255	0.5975	0.7451	0.2623
(f)	√		√					√		0.6430	0.8241	0.6036	0.7485	0.6110	0.7863	0.2551
(e)	√	√							√	0.6387	0.7969	0.5924	0.7322	0.6027	0.7739	0.2568
(g)	√		√						√	0.6680	0.8325	0.6365	0.7504	0.6278	0.7885	0.2421
(h)	√		√		√				√	0.6938	0.8605	0.6417	0.7604	0.6290	0.7883	0.2396
(i)	√		√			√			√	0.7051	0.8733	0.6696	0.7595	0.6393	0.7952	0.2388
(j)	√			√		√			√	0.7289	0.8923	0.6990	0.7834	0.6571	0.8120	0.2221
(k)	√		√			√	√		√	0.7190	0.8912	0.6824	0.7677	0.6495	0.8052	0.2294
(l)	√			√		√	√		√	**0.7433**	**0.9030**	**0.7075**	**0.7994**	**0.6755**	**0.8250**	**0.2106**

TABLE III
ABLATION EXPERIMENTS DESCRIBING THE WAY OF SETTING β. COMPARED WITH SETTING THE VALUE OF β MANUALLY, THE DYNAMIC VALUE OF β FURTHER IMPROVES THE PERFORMANCE.

β	0	0.2	0.4	0.6	0.8	ours
mAP_8	0.7303	0.7321	0.7320	0.7347	0.7266	0.7433
mAP_2	0.8870	0.8885	0.8923	0.8972	0.8867	0.9030

performance. Taking into consideration both the performance and computational consumption, we extract the low-level and high-level features from the last layer of conv_3 and conv_5, respectively. The semi-hard triplet sampling method is applied in the triplet loss to guarantee the model converges stably and rapidly. In the baseline models, the feature vector is obtained through the global average pooling operation on the feature map from the last convolutional layer. The dimension of output feature embedding is compacted to 512 following the empirical insights in [75]. We randomly split the FI dataset into 80% training, 5% validation, and 15% test sets. The parameters of the model trained on FI are transferred to fine-tune the other small-scale datasets. We conduct 5-fold validation and report the average performance. The entire work is implemented using PyTorch, where all experiments are conducted on one NVIDIA GTX 1080Ti GPU.

E. Retrieval Performance

The effectiveness of the proposed method is validated on four emotional datasets. In Tab. I, we report the results of various contrastive methods of attention networks and deep metric learning on the FI dataset. It is obvious that the end-to-end learning-based methods perform better than those based on handcrafted features, such as SIFT, HOG, and Gabor. SentiBank, DeepSentiBank and MVSO belong to the same series of algorithms that can detect the ANP concepts for each image as the mid-level representations. Among the three types of representations, the performance is slightly improved with the increase of feature dimensions. Generally, the network optimized by metric loss achieves remarkably better overall performance than those with the supervision of softmax loss. Note that the performance of metric learning on the 'NN' metric cannot outperform that of the softmax loss as on other metrics. This is because the softmax loss mainly concerns the boundary between different categories but ignores the concrete distance between feature embeddings. Meanwhile, the metric loss directly manipulates features in the embedding space to maximize the inter-class variation and minimize the intra-class variation. Therefore, the feature points learned by metric loss can well distribute in the embedding space according to the emotion similarity.

Furthermore, we also compare the proposed method with the latest and popular metric learning algorithms as well as state-of-the-art methods [90] for emotion-based image retrieval. Particularly, to achieve a fair comparison, we implement the state-of-the-art algorithms using ResNet-50 as the backbone, which is the same as that in our method. Obviously, our framework achieves much better performance than the state-of-the-art methods, especially on mAP_2 and mAP_8 (approximately 10% improvement). Compared with the results of the conference version paper, the methods of generating negative embeddings utilized in this journal paper further improve the retrieval performance on six of seven metrics.

For the three small-scale datasets, we directly fine-tune the network using the training dataset based on the model that has been trained on the FI dataset. These datasets include natural images and abstract art images in which there is a large domain gap. As shown in Fig. 5, our method also obtains the best retrieval results, which demonstrates the robust generalization ability of our method for different domains.

F. Ablation Study

To present an in-depth analysis of the effect of each component in the proposed framework, we conduct a detailed ablation study and show the experimental results on the FI dataset in Tab. II. In the first part, we verify the effectiveness of EP loss and the features at multiple levels. First, AT

TABLE IV
RESULTS OF DIFFERENT COMBINATION STRATEGIES AMONG CONVOLUTIONAL LAYERS. 'P' DENOTES POLARITY-SPECIFIC ATTENDED FEATURES, WHILE 'E' REPRESENTS EMOTION-SPECIFIC ATTENDED FEATURES. SINCE THE COMBINATION OF CONV_3 AND CONV_5 PERFORMS BEST ON SIX OUT OF SEVEN CRITERIA, WE EMPLOY THIS STRATEGY IN ALL EXPERIMENTS.

combinations	mAP_8 ↑	mAP_2 ↑	FT↑	ST↑	NN↑	DCG↑	ANMRR↓
conv_2(p)+conv_5(e)	0.7351	0.8989	0.6890	0.7912	0.6589	0.8152	0.2209
conv_2(e)+conv_5(e)	0.7304	0.8901	0.6934	0.7881	0.6623	0.8136	0.2253
conv_3(p)+conv_5(e)	**0.7433**	**0.9030**	**0.7075**	**0.7994**	**0.6755**	0.8250	**0.2106**
conv_3(e)+conv_5(e)	0.7352	0.8928	0.6951	0.7892	0.6661	0.8179	0.2191
conv_4(p)+conv_5(e)	0.7335	0.9012	0.6982	0.7912	0.6682	0.8185	0.2146
conv_4(e)+conv_5(e)	0.7356	0.8981	0.6868	0.7739	0.6622	0.8046	0.2130
conv_5(p)+conv_5(e)	0.7316	0.8969	0.6922	0.7877	0.6626	0.8069	0.2250
conv_5(e)+conv_5(e)	0.7380	0.8912	0.7012	0.7920	0.6678	**0.8271**	0.2163

represents the attention loss conducted on conv_3 and conv_5, where the attention loss includes two softmax losses. When N-pair loss serves as the optimization function, the performance is obviously improved compared with that based on attention loss. It is mainly because N-pair loss can directly manipulate the distance between different feature embeddings. Further, the results of the proposed EP loss outperform N-pair loss on all the metrics, especially on mAP_2, which demonstrates that the EP loss well learns the decision boundary between polarities. Meanwhile, the improvement on mAP_2 also facilitates an approximate 9% increase on mAP_8. Obviously, benefiting from the mutual promotion of multiple tasks, simultaneously exploiting AT and EP losses can obtain better performance on all the metrics except mAP_2. The reduction on mAP_2 is mainly because the AT of the last convolution layer ignores the boundary between two polarities. The slight reduction will be recovered by the multilevel outputs (shown in (e) and (g)) and attention mechanism.

Furthermore, we also ablate how to design the attention module to obtain better performance. The detailed experimental results are shown in the second part of Tab. II. By incorporating emotion-specific attention into conv_5, the performances on mAP_2 and mAP_8 gain 4% and 3% improvements, respectively. It indicates that some informative regions of the image can actually provide more abundant emotional features. When both polarity and emotion-specific attention modules are utilized in our framework, the results are further improved, which demonstrates that the attended features from different levels capture more useful information. To make the multilevel features interact effectively, CLB (shown in (k) and (l)) is introduced to obtain higher-order information, leading to further improvement over the baseline that directly concatenates them. Finally, the proposed method of generating sample pairs adaptively (*i.e.*, GEP loss) improves the overall performance effectively.

In Tab. III, we ablate the way of setting the value of β. With a fixed value of β, λ_0 will be only determined by the ratio between the distance of the positive pair and the distance of the negative pair, *i.e.*, $\frac{d(f_i, f_i^+)}{d(f_i, f_j^+)}$. In our method, the value of β is adaptively controlled based on the hardness of each sample. Therefore, all the training examples are fully utilized, resulting in a more discriminative model.

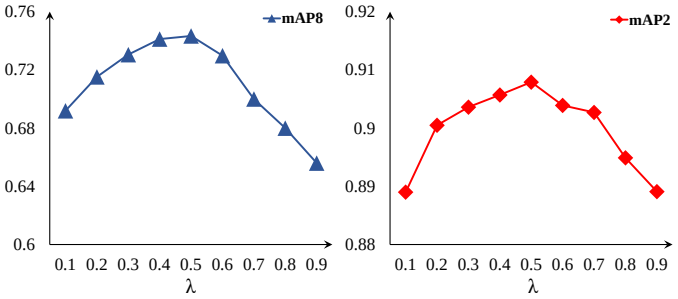


Fig. 6. Effect of λ for the total loss on mAP_8 and mAP_2 testing on the FI dataset. Note that λ is the weight of $\mathcal{L}_{gep}$, and $1-\lambda$ is the weight of $\mathcal{L}_{att}$.

G. Combinations of Multiple Stages

In Tab. IV, we discuss the combinations among four stages (conv_2, conv_3, conv_4, and conv_5) in ResNet-50 and only extract the feature maps from the last layer in each stage. As shown in Tab. IV, the combination of conv_3(p) and conv_5(e) performs the best on six out of seven criteria, where p means the polarity-specific attended features and e means the emotion-specific attended features. On the one hand, the features from conv_3 and conv_5 interact better than other combinations. On the other hand, the attended regions relevant to sentiment polarity from conv_3 provide significant complementary cues with high-level features. Therefore, we select the combination of conv_3 and conv_5 in all the experiments.

H. Influence of Parameter λ

Based on the FI dataset, we discuss the sensitivity of hyper-parameter λ, which controls the relative importance between the GEP loss and attention loss in Eq. (14). In Fig. 6, the results on mAP_8 and mAP_2 are shown when λ ranges from 0.1 to 0.9. We can draw two conclusions from the curves: (1) mAP_8 is more sensitive than mAP_2 for the variation of λ; and (2) when $\lambda = 0.5$, mAP_8 and mAP_2 both achieve the best performance. Note that the performance on mAP_8 descends dramatically when $\lambda > 0.6$, which means the weight of attention loss is less than 0.4. It is concluded that softmax loss (attention loss) can guide metric loss to recognize the concrete categories. Then, the metric loss can well manipulate the Euclidean distance between features.

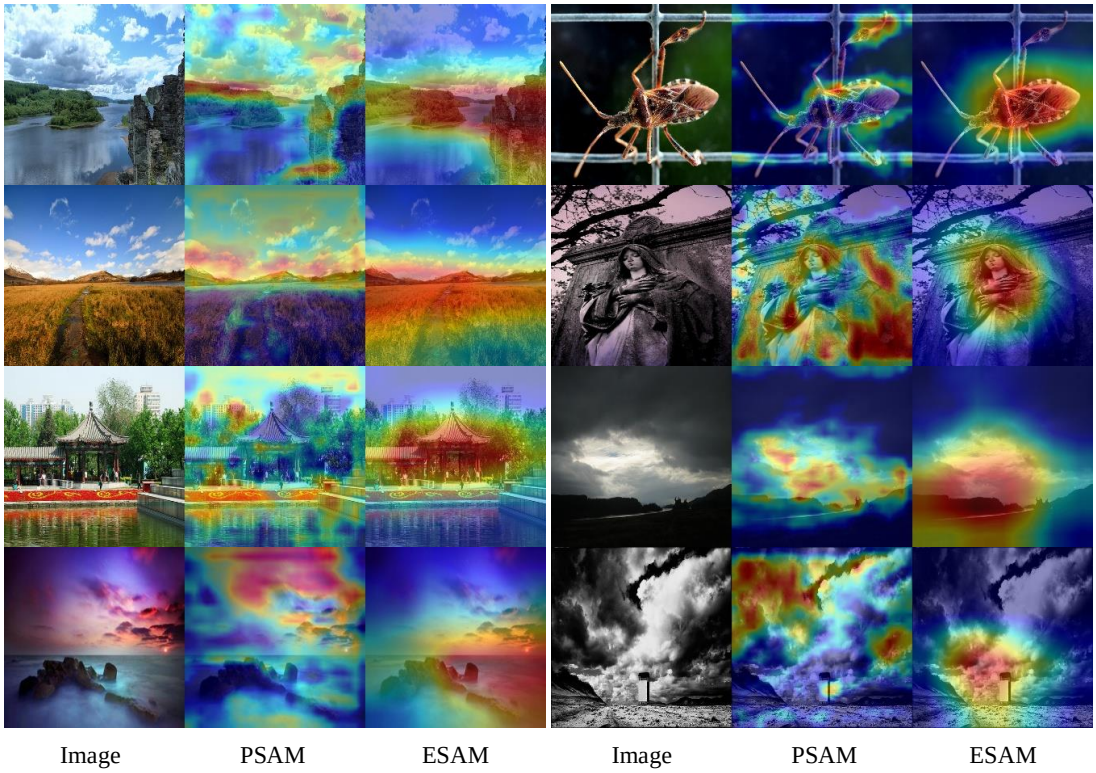


Fig. 7. Visualization of attention maps of different levels. For each image from the FI dataset, we show its corresponding polarity-specific attention map (PSAM) and emotion-specific attention map (ESAM).

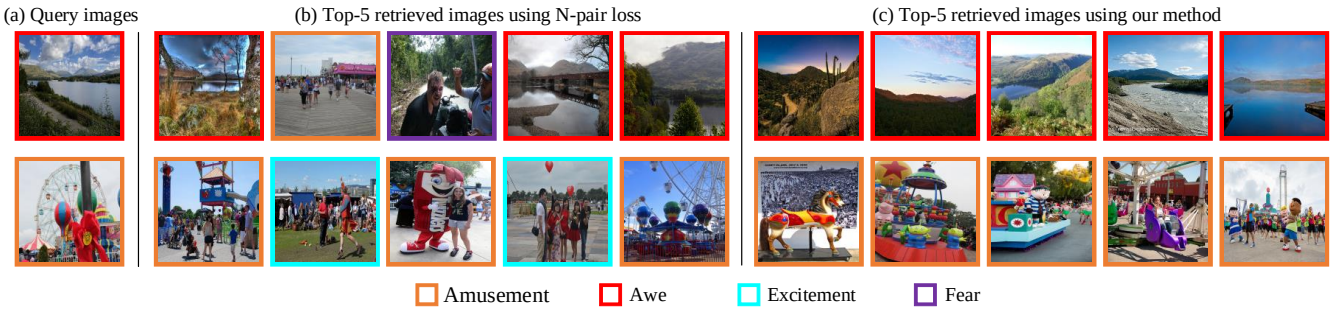


Fig. 8. Top-5 results of example query images from the FI dataset. (a) are query images from FI. (b-c) are the retrieval results of networks trained by the N-pair loss and our method, respectively. Image frames with different colors represent different emotions.

I. Visualization

We randomly select several attentional visualization results in Fig. 7. For the first and the third instances of the right column, some regions with distinct color and texture are highlighted by polarity-specific attention. They can be regarded as cues to guide the emotion-specific attention to be more concerned with complete regions. However, for the second image of the right column, polarity-specific attention mainly focuses on the gloomy surroundings of the statue, while emotion-specific attention is more concerned about the face of statue in which high-level emotional semantic information can be conveyed. The polarity-specific attended region can well complement the emotion-specific attended region.

In Fig. 8, we present the top-5 retrieved images from the FI dataset learned by N-pair loss and our method. With the supervision of N-pair loss, even images from the opposite polarity appear in the top-5 results, such as the results for the first query. This is due to the negligence of local information (*e.g.*, the big spider in the man's face of the third returned image) and the hierarchy of emotion. By contrast, the proposed method obtains the correct results in the top-5 images for the two examples.

In Fig. 9, we show some failure cases of our method. For the first query of excitement, there are two images of awe in the top-3 results. In fact, the two images can also make viewers feel excited, which is due to the emotional diversity of one image. That is, the emotional boundary of some images is ambiguous. The disgust emotion of the second query is caused by the content of the magazine on the desk, which is difficult for us to see clearly. Therefore, these types of failure cases may be lessened by improving the resolution of images.

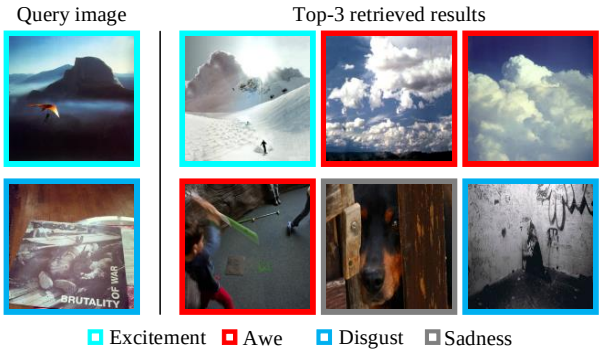


Fig. 9. Representative failure cases in top-3 results. The first example is from the IAPSa dataset, and the second example is from the FI dataset.

V. Conclusion and Future Work

In this paper, we proposed an APSE network for emotion-based image retrieval. In the hierarchical attention module, the polarity- and emotion-specific attended features are effectively integrated through the cross-level bilinear operation. We developed a GEP loss for feature embedding learning, which constrains features from inter- and intra-polarity simultaneously. The negative examples can be generated adaptively based on confidence scores derived from the attention module. Finally, multiple losses, including GEP and attention losses, are employed to optimize the framework. Extensive experiments on four datasets demonstrate that the proposed framework outperforms the state-of-the-art approaches.

For further studies, we will try to take into account the ambiguity of emotion for EBIR. For example, the similarity between emotional images can be measured by the distances between the label distributions of images. In addition to discrete label space, retrieving emotional images in continuous label space, such as valence-arousal space, is also a meaningful topic for some professional applications.

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.20JCJQJC00020, 18JCYBJC15400, 18ZXZNGX00110), and the Fundamental Research Funds for the Central Universities.

References

[1] X. Yao, D. She, S. Zhao, J. Liang, Y.-K. Lai, and J. Yang, "Attention-aware polarity sensitive embedding for affective image retrieval," in *ICCV*, 2019.

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

[3] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz, "Emotional category data on images from the international affective picture system," *Behavior Research Methods*, vol. 37, no. 4, pp. 626–630, 2005.

[4] H. Zhang and M. Xu, "Recognition of emotions in user-generated videos with kernelized features," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2824–2835, 2018.

[5] G. Tu, Y. Fu, B. Li, J. Gao, Y.-G. Jiang, and X. Xue, "A multi-task neural approach for emotion attribution, classification, and summarization," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 148–159, 2019.

[6] K. Ahmad, S. Zohaib, N. Conci, and A. Al-Fuqaha, "Deriving emotions and sentiments from visual content: A disaster analysis use case," *arXiv preprint arXiv:2002.03773*, 2020.

[7] X. Guo, L. Polania, B. Zhu, C. Boncelet, and K. Barner, "Graph neural networks for image understanding based on multiple cues: Group emotion recognition and event recognition as use cases," in *WACV*, 2020.

[8] F. Chen, R. Ji, J. Su, D. Cao, and Y. Gao, "Predicting microblog sentiments via weakly supervised multimodal deep learning," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 997–1007, 2017.

[9] S. C. Guntuku, D. Preotiuc-Pietro, J. C. Eichstaedt, and L. H. Ungar, "What twitter profile and posted images reveal about depression and anxiety," in *AAAI*, 2019.

[10] H. Lin, J. Jia, Q. Guo, Y. Xue, J. Huang, L. Cai, and L. Feng, "Psychological stress detection from cross-media microblog data using deep sparse neural network," in *ICME*, 2014.

[11] S. Pan, J. Lee, and H. Tsai, "Travel photos: Motivations, image dimensions, and affective qualities of places," *Tourism Management*, vol. 40, pp. 59–69, 2014.

[12] Q. You, H. Jin, and J. Luo, "Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions," *arXiv preprint arXiv:1801.10121*, 2018.

[13] B. Jou, S. Bhattacharya, and S.-F. Chang, "Predicting viewer perceived emotions in animated gifs," in *ACM MM*, 2014.

[14] Z. Yang, Y. Zhang, and J. Luo, "Human-centered emotion recognition in animated gifs," in *ICME*, 2019.

[15] 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.

[16] W. Wei-ning, Y. Ying-lin, and J. Sheng-ming, "Image retrieval by emotional semantics: A study of emotional space and feature extraction," in *SMC*, 2006.

[17] Y. Kim, Y. Shin, S.-j. Kim, E. Y. Kim, and H. Shin, "EBIR: Emotion-based image retrieval," in *ICCE*, 2009.

[18] 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.

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

[20] 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.

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

[22] J. Wang, J. Fu, Y. Xu, and T. Mei, "Beyond object recognition: Visual sentiment analysis with deep coupled adjective and noun neural networks," in *IJCAI*, 2016.

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

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

[25] M. G. Calvo and P. J. Lang, "Gaze patterns when looking at emotional pictures: Motivationally biased attention," *Motivation and Emotion*, vol. 28, no. 3, pp. 221–243, 2004.

[26] 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.

[27] 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.

[28] 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.

[29] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *NIPS*, 2016.

[30] 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.

[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] D. She, J. Yang, M.-M. Cheng, Y.-K. Lai, P. L. Rosin, and L. Wang, "Wscnet: Weakly supervised coupled networks for visual sentiment classification and detection," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1358 – 1371, 2020.

[33] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *CVPR*, 2018.

[34] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2018.

[35] W. Ge, W. Huang, D. Dong, and M. R. Scott, "Deep metric learning with hierarchical triplet loss," in *ECCV*, 2018.

[36] X. Wang, Y. Hua, E. Kodirov, G. Hu, and N. M. Robertson, "Deep metric learning by online soft mining and class-aware attention," *arXiv preprint arXiv:1811.01459*, 2018.

[37] 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.

[38] C. Lin, S. Zhao, L. Meng, and T.-S. Chua, "Multi-source domain adaptation for visual sentiment classification," in *AAAI*, 2020.

[39] T. Liu, J. Wan, X. Dai, F. Liu, Q. You, and J. Luo, "Sentiment recognition for short annotated gifs using visual-textual fusion," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 1098–1110, 2020.

[40] J. Yuan, S. Mcdonough, Q. You, and J. Luo, "Sentribute: image sentiment analysis from a mid-level perspective," in *ACM International Workshop on Issues of Sentiment Discovery and Opinion Mining*, 2013.

[41] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *ACM MM*, 2013.

[42] P. Yang, Q. Liu, and D. N. Metaxas, "Exploring facial expressions with compositional features," in *CVPR*, 2010.

[43] D. She, M. Sun, and J. Yang, "Learning discriminative sentiment representation from strongly-and weakly supervised cnns," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 3s, pp. 1–19, 2019.

[44] 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.

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

[46] 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.

[47] L. Li, X. Zhu, Y. Hao, S. Wang, X. Gao, and Q. Huang, "A hierarchical cnn-rnn approach for visual emotion classification," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 3s, pp. 1–17, 2019.

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

[49] K. A. Olkiewicz and U. Markowska-Kaczmar, "Emotion-based image retrievalan artificial neural network approach," in *IMCSIT*, 2010.

[50] B. Xing, K. Zhang, S. Sun, L. Zhang, Z. Gao, J. Wang, and S. Chen, "Emotion-driven chinese folk music-image retrieval based on de-svm," *Neurocomputing*, vol. 148, pp. 619–627, 2015.

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

[52] T. Chen, Z. Zhang, Q. You, C. Fang, Z. Wang, H. Jin, and J. Luo, "Factual or emotional: Stylized image captioning with adaptive learning and attention," in *ECCV*, 2018.

[53] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *ICCV*, 2019.

[54] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *CVPR*, 2018.

[55] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *ICCV*, 2019.

[56] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *CVPR*, 2017.

[57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[58] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018.

[59] P. Vuilleumier, "How brains beware: neural mechanisms of emotional attention," *Trends in Cognitive Sciences*, vol. 9, no. 12, pp. 585–594, 2005.

[60] R. J. Compton, "The interface between emotion and attention: A review of evidence from psychology and neuroscience," *Behavioral and Cognitive Neuroscience Reviews*, vol. 2, no. 2, pp. 115–129, 2003.

[61] 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.

[62] 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.

[63] Q. You, L. Cao, H. Jin, and J. Luo, "Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks," in *ACM MM*, 2016.

[64] S. Kim, M. Seo, I. Laptev, M. Cho, and S. Kwak, "Deep metric learning beyond binary supervision," in *CVPR*, 2019.

[65] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *CVPR*, 2017.

[66] V. E. Liong, J. Lu, Y.-P. Tan, and J. Zhou, "Deep coupled metric learning for cross-modal matching," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1234–1244, 2016.

[67] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *CVPR*, 2005.

[68] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *CVPR*, 2015.

[69] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *CVPR*, 2016.

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

[71] W. Zheng, Z. Chen, J. Lu, and J. Zhou, "Hardness-aware deep metric learning," in *CVPR*, 2019.

[72] L. Liao, X. He, B. Zhao, C.-W. Ngo, and T.-S. Chua, "Interpretable multimodal retrieval for fashion products," in *ACM MM*, 2018.

[73] D. Wang, H. Huang, C. Lu, B.-S. Feng, L. Nie, G. Wen, and X.-L. Mao, "Supervised deep hashing for hierarchical labeled data," in *AAAI*, 2018.

[74] Y. Peng, J. Qi, X. Huang, and Y. Yuan, "Ccl: Cross-modal correlation learning with multigrained fusion by hierarchical network," *IEEE Transactions on Multimedia*, vol. 20, no. 2, pp. 405–420, 2017.

[75] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *CVPR*, 2015.

[76] D. G. Lowe, "Object recognition from local scale-invariant features," in *ICCV*, 1999.

[77] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

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

[79] 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.

[80] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

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

[82] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.

[83] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[84] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, 2016.

[85] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *ICPR*, 2014.

[86] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *CVPR*, 2017.

[87] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019.

[88] F. Cakir, K. He, X. Xia, B. Kulis, and S. Sclaroff, "Deep metric learning to rank," in *CVPR*, 2019.

[89] Q. Qian, L. Shang, B. Sun, J. Hu, H. Li, and R. Jin, "Softtriple loss: Deep metric learning without triplet sampling," in *ICCV*, 2019.

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

[91] P. J. Lang, M. M. Bradley, B. N. Cuthbert *et al.*, "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.

[92] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422–446, 2002.

[93] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai, "3-D object retrieval and recognition with hypergraph analysis," *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 4290–4303, 2012.