Applied Intelligence (2021) 52:10901–10916
https://doi.org/10.1007/s10489-021-02960-9

Three-way enhanced part-aware network for fine-grained sketch-based image retrieval

Xiuying Wang[1] · Jun Tang[1] · Shoubiao Tan[1]

Accepted: 26 October 2021 / Published online: 18 January 2022

Abstract

Sketch-based image retrieval is of import practical significance in today's world populated by smart touch screen devices. Fine-grained sketch-based image retrieval (FG-SBIR) is particularly challenging and uses characteristic of free-hand sketches to retrieve natural photos at the instance level. From outline and semantic perspectives, a free-hand sketch may have many natural photos corresponding to it, we call the relationship "one-to-many", which means that the effectiveness of FG-SBIR mainly depends on the quality of fine-grained information extracted. Existing deep convolutional neural network (DCNN) models for FG-SBIR commonly use coarse or first-order attention modules to focus on specific local regions, yet cannot capture high-order or complex information and the subtle differences between sketch–photo pairs. It is widely known that the features learned from higher layers of the network are more abstract and of a higher semantic level compared to those learned from the lower layers, but lose some important fine-grained information. To address these limitations, this paper proposes a three-way enhanced part-aware network (EPAN), in which a mixed high-order attention module is added after the middle-level feature space to generate a variety of high-order attention maps and capture rich features contained in the middle convolutional layer. An enhanced part-aware module is proposed to capture useful part cues and enhance the semantic consistency of local regions. This allows for learning more discriminative cross-domain feature representations. A larger number of experiments on several popular datasets demonstrate that our model is superior to state-of-the-art approaches.

Keywords Sketch-based image retrieval · Fine-grained · Enhanced part-aware network · High-order attention

1 Introduction

The use of image retrieval has increased in social media, e-commerce, medicine, and other fields with the prevalence of the multimedia data on the internet. Two approaches are commonly used for image retrieval: text-based image retrieval (TBIR) [1] and content-based image retrieval (CBIR) [2, 3]. The former searches for an image by providing a text description and the latter by providing a similar image as a query image. In some cases, it may be difficult to provide an accurate textual description for the query image, and it may not always be possible to give a real image as a query. Sketches are easy to draw and usually convey richer and more compact information than text in some scenarios. Therefore, sketch-based image retrieval (SBIR) has drawn considerable attention recently [4–15].

The goal of SBIR is to retrieve related photos from a database for the specific query sketch. Research on SBIR mainly occurs in two fronts: category-level sketch-based image retrieval(c-SBIR) [11, 12, 14–17] and fine-grained sketch-based image retrieval (FG-SBIR) [4, 5, 7, 9, 18]. As shown in Fig. 1, the goal of c-SBIR is to find natural photos of the same category for a query sketch, while the goal of FG-SBIR is to find a unique corresponding natural photo for the query sketch. The key difference between them is the granularity of the retrieval results. Compared with c-SBIR, the FG-SBIR fully exploits the details that can be conveyed in sketches. This article is interested in the more challenging FG-SBIR task.

✉ Shoubiao Tan
tsb@ustc.edu

Xiuying Wang
wxy_ahu@163.com

Jun Tang
tangjunahu@163.com

[1] School of Electronics and Information Engineering, Anhui University, Hefei 230601, China

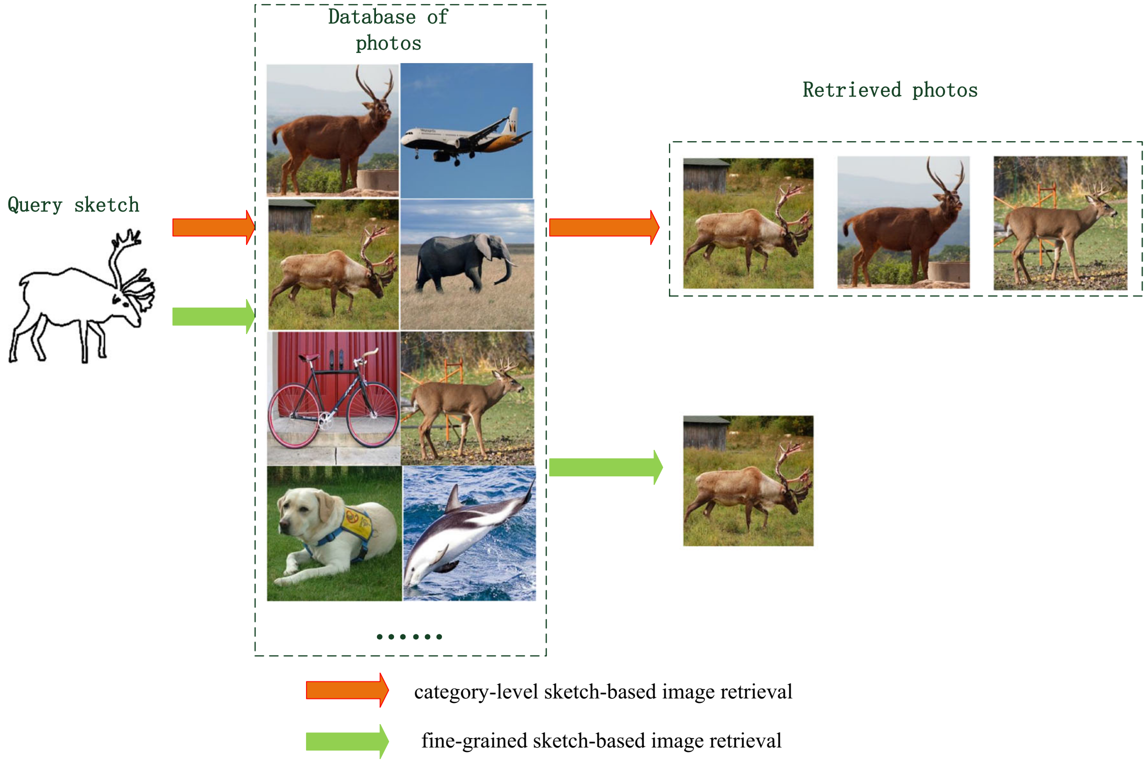

Database of photos
Retrieved photos
Query sketch
......
category-level sketch-based image retrieval
fine-grained sketch-based image retrieval

Fig. 1 Illustration of Sketch based Image Retrieval

Two major challenges for FG-SBIR are illustrated in Fig. 2: (1) sketches and photos come from two domains, exhibiting a large domain gap. The sketch is abstract, mainly composed of shape and contour information, while the photo contains rich color and texture information. There are some differences in spatial position and contour between the sketch and the matched photo. (2) From an outline and semantics perspective, a query sketch may have many similar natural photos in the database. Similarities between different candidate photos and imperfect detail descriptions further increase the difficulty of retrieval. To solve the first problem, recently, deep feature learning-based methods [4, 5, 7, 8, 19–21] have used DCNN to learn a joint embedding representation. They usually used triplet loss or contrastive loss for cross-domain similarity learning. There exist several methods [4, 5, 22] that adopt edge maps as the intermediate representations to bridge the gap between free-hand sketches and natural photos. Meanwhile, several researchers [23–26] have attempted to bridge the cross-domain gap using generative adversarial networks (GANs) [27], which translate a natural photo (or sketch) into its corresponding sketch (natural photo). To solve the second problem, many studies [4, 16] introduced an attention mechanism in the network, with the purpose of focusing representation learning on specific discriminative local regions. Some studies [4, 28] integrated coarse and fine features through a fusion module to obtain different levels of semantic representation.

Although these methods have obtained impressive results, several difficulties and challenges remain. 1) The structure of existing DCNN models for FG-SBIR is mostly composed of several convolution and fully connected layers. The joint embedding features they learned for different domains are usually global, without considering the importance of part-level features for fine-grained retrieval; moreover, they do not take full advantage of the rich features involved in the middle convolutional layer, and the expression ability of the learned feature vector was insufficient. 2) The common attention methods are coarse-grained and first-order, such as spatial attention and channel attention. They have limited ability to obtain complex high-order information and cannot effectively capture subtle differences between sketch–image pairs. 3) Most widely used loss function of deep SBIR methods is triplet loss [29]. The original triplet loss function only requires that the feature distances between inter-class to be larger than that of the intra-class ones, and ignoring the compactness of intra-class (as shown in Fig. 3).

To address the above limitations, we propose a new architecture called three-way enhanced part-aware network in this paper. In FG-SBIR benchmark datasets, the fine-grained sub-categories contain few samples, edge maps

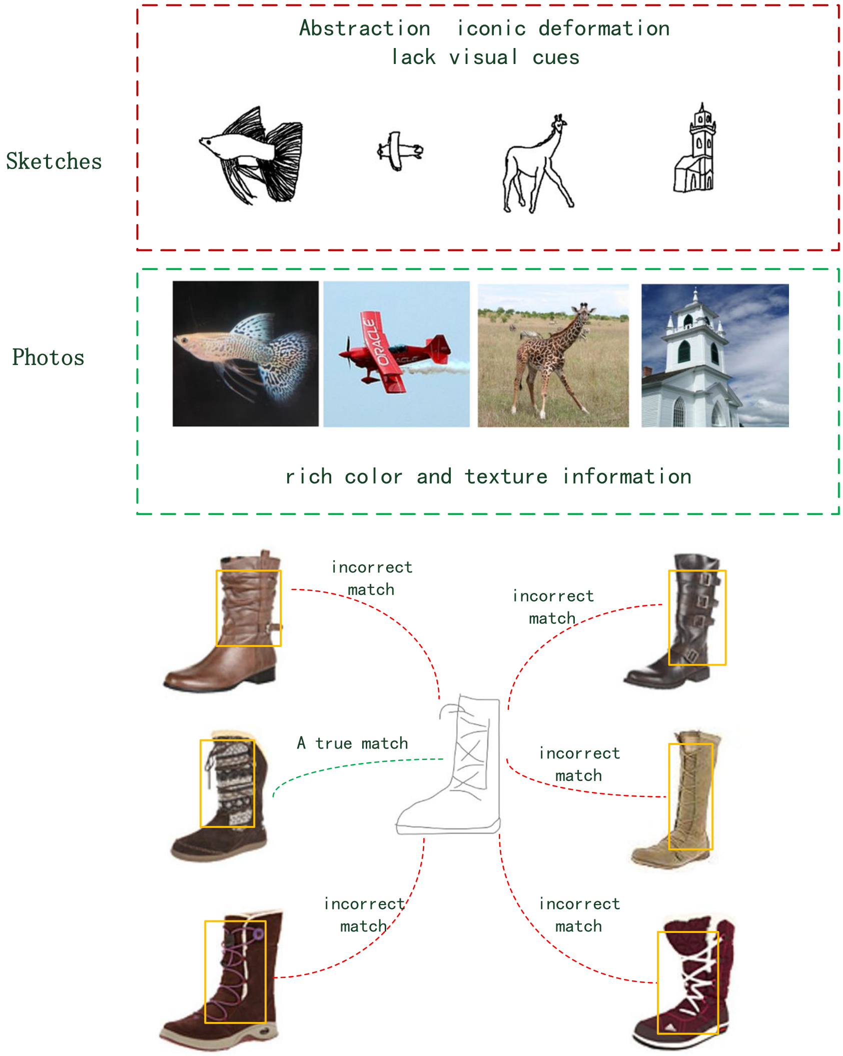


Fig. 2 An illustration of FG-SBIR challenges

can increase the semantic information available about sub-categories. Moreover, from the perspective of visual abstraction, edge maps are closer to sketches than natural photos. Therefore, enabling full use of the auxiliary role of edge maps, we design a three-way network consists of natural photo branch, sketch branch and edge map branch. In each branch network, the stacked convolutional layers are first decomposed into two parts: P1 and P2. P1 and P2 represent the mid-level and high-level encode network, respectively. We place the mixed high-order attention module [30] between P1 and P2 to generate diverse high-order attention maps to capture rich features

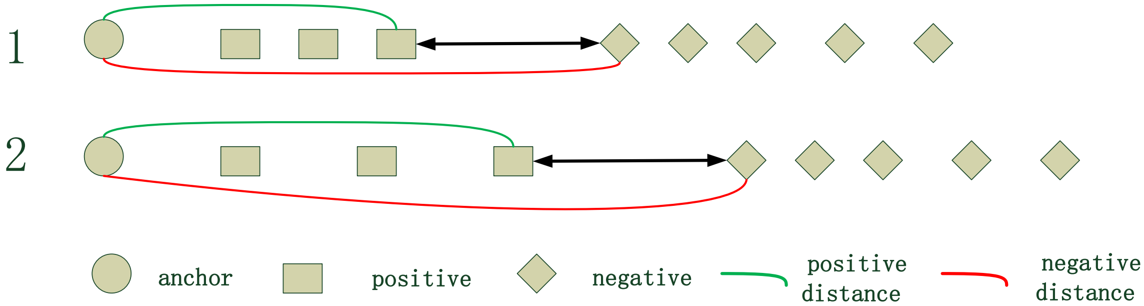


Fig. 3 Illustration of original triplet loss, which ignores intra-class compactness of positive pairs. Two cases can acquire the same inter-class separability, but the first case has better intra-class compactness, making it easier to choose an appropriate threshold to separate positive and negative samples

contained in the middle convolutional layer and produce the discriminative attention proposals. Part-level features offer fine-grained information and have been verified to be beneficial for image retrieval recently, thus, an enhanced part-aware module is designed to capture useful part cues and integrate global structural information for an input feature, where a self-attention module is adopted to learn the relationship among different part features. In addition, to increase embedding discriminability, we propose a hybrid-loss mechanism, which consists of adversarial loss, an improved bidirectional triplet loss and classification loss. An adversarial learning constraint is applied to the mixed high-order attention module to prevent the order collapse problem, an improved bidirectional triplet loss function to enhance the separability of inter-class and the compactness of intra-class. Furthermore, the classification losses are applied to reinforce semantic consistency of local parts.

Our contributions are: (1) a mixed high-order attention module is added after the middle level feature space of each branch network, and the higher-order relationships among parts are established to generate powerful attention proposal. In the FG-SBIR task, the ability to obtain discriminative fine-grained information is improved, which can be used as a reference for other fine-grained research. (2) We fully consider the importance of part-level features to fine-grained retrieval and combine self-attention to design an enhanced part-aware module to learn the relationship among different part features and improve the performance of cross-domain retrieval. (3) A hybrid-loss mechanism is designed to reduce the domain gap and facilitate the network learning for FG-SBIR task. Experimental results show that the superiority of our model over the state-of-the-art approaches on three FG-SBIR popular datasets.

2 Related work

2.1 Deep Fine-grained SBIR

Compared with c-SBIR, FG-SBIR has more potential for real-world application. The FG-SBIR problem was first studied in [31] using graph matching on deformable part-based models for sketches and photographs. Subsequently, more and more DNN-based networks have been proposed to tackle the challenges of FG-SBIR [4, 5, 7, 28, 32]. Yu et al. [5] proposed a deep triplet-ranking model to train a joint embedding representation for sketch domain and edge maps domain. Then, this model was subsequently improved by introducing coarse-fine fusion, attention modelling, and HOLEF loss [4]. Yu et al. [28] adopted a fusion module to obtain mid-level and high-level combined features. TC-Net [32] added an auxiliary classification loss to a triplet Siamese network to provide an intra-class constraint for paired photos and sketches. Many methods [23, 24, 33] have been exploited from an image synthesis perspective and have achieved better performance. Inspired by the recent BERT model [34] for natural language processing, [35] and [36] have been exploited and significantly improved the effect of sketch retrieval. There are also some new research directions of FG-SBIR. For example, the cross-category FG-SBIR generalization (CC-FG-SBIR) problem was first proposed and solved in [37]. Bhunia et al. [38] studied a new FG-SBIR network to tackle the problem for on-the-fly and early sketch retrieval.

2.2 Attention mechanism

Most recent researches on the attention mechanism of deep learning have focused on using masks to form attention and made the model biased to the region with the most abundant information. The mask re-weights the input feature map through a new layer to obtain an attended feature map and feed it to the next layer of the network. Attention can be grouped in terms of calculation regions: hard attention [40] and soft attention [39]. Soft-attention mechanism focus on the global, which assigns different weights to each pixel of the source image while hard-attention mechanism is a random process that only focus on a small region at each time. For instance, Li et al. [41] designed a hard attention mechanism locates latent regions to extract and exploited these regional features for ReID. A channel attention model was proposed in [42] to assign different weights to each channel, which effectively improves the classification performance. A soft attention spatial model was proposed in [4], which use a weight mask to reweight the different spatial regions of the feature map. This model effectively enhances the ability to discriminate fine-grained features and achieves good retrieval performance. However, the attention mechanisms for fine-grained SBIR task are coarse and do not have the ability to acquire high-order relationships among different regions. To address these, a mixed high-order attention module [30] is adopt in this paper for FG-SBIR task to capture more fine-grained information. Most previous FG-SBIR models typically added the attention modules after the last convolutional layer. However, such models cannot take full advantage of the rich features contained in the middle convolutional layer. Thus, we divide the convolutional layer of the base network into two parts and add this attention module between them.

2.3 Loss function

The loss functions extensively used in the field of FG-SBIR are triplet loss, contrastive loss, and their variants.

Triplet loss The basic target of triplet loss is that the distance between the negative pairs should be larger than the positive pairs by a pre-defined margin [29]. For a given triplet $t = (s, p^+, p^-)$, s, p^+, p^- denote a sketch, a positive photo, and a negative photo respectively. A conventional triplet loss is computed as:

$$L_{tri}(s, p) = [D(f^t(s), f^t(p^+)) - D(f^t(s), f^t(p^-)) + \varDelta]_+ \tag{1}$$

where $f^t(\cdot)$ denotes the feature function. $D(\cdot)$ denotes a distance between two features, $\varDelta$ is a margin parameter and $[\theta]_+ = max(\theta; 0)$. The $D(\cdot)$ can be represented by different distance functions, such as Euclidean distance [5, 32] and a higher-order energy function [4].

Triplet loss is greatly affected by the sample triplets: if the training triplets contain a lot of easy triplets, the discriminability of the model will be limited; Using an appropriate distance threshold is critical for triplet loss: if the margin value is set too small, the loss can approach 0, making it difficult to distinguish between similar images. Otherwise, the network will not converge. To address these, various methods have been proposed [43–45].

Contrastive loss The training input pairs consists of sketch s_i and photo p_j, and define the label $Y(i, j)$, for which the value is zero or one. When p_j is the positive sample of sketch s_i, the label $Y(i, j) = 1$. Conversely, if p_j is a negative sample, $Y(i, j) = 0$. The contrastive loss [46] that acts on non-matching and matching sketch-photo pairs to further constrain the relationship between them is defined as:

$$\begin{aligned} L(i, j) = &\frac{1}{2} Y(i, j) \cdot D(f^t(s_i), f^t(p_j)) \\ &+ \frac{1}{2}(1 - Y(i, j)) \cdot \max(0, \varDelta^1 - D(f^t(s_i), f^t(p_j))) \end{aligned} \tag{2}$$

where $\varDelta^1$ is a margin parameter.

Combination of classification loss and triplet loss can improve the feature representation ability of the FG-SBIR model. Lin et at. [32] added an auxiliary classification loss to force the matched pairs closer to each other in the embedding space. In this paper, we also use classification loss to reinforce the semantic consistency of local parts.

3 Proposed method

3.1 Overall framework

Free-hand sketches are usually abstract, monotonous, and ambiguous. The processing efficiency of sketches depends not only on the global structure but also on the local details. For example, when the contour of the input sketch is circular, we can classify it roughly (e.g., as the moon, an alarm clock, a hot air balloon), but more details are needed for accurate classification. We propose a new three-way enhanced part-aware network is illustrated in Fig. 4. It consists of the sketch branch FB_{st}, the natural photo branchFB_{im}, and the edgemap branch FB_{em}, the weights of FB_{im}, FB_{st} and FB_{em} are completely share. Each branch includes the following distinct parts: a CNN base network (blue box), a mixed high-order attention module (orange box), a part-aware module (red box), a semantic embedding module (green box) and a hybrid loss module (black box). Specifically, The Alexnet architecture without the fully connected layers is adopted by us as the base network. All the convolutional layers are decomposed into two parts: P1 (from conv1 to conv2) and P2 (from conv3 to conv5). These parts are used to encode the input feature to mid-level or high-level feature space. A mixed high-order attention module is constituted by four different high-order attention (HOA) modules, which placed between P1 and P2 to capture rich features contained in the middle convolutional layer and produce the diverse high-order attention maps. Given a high-level feature, a part-aware module can produce part-level features that contain more information that is detailed. Then, a self-attention module can capture enhanced feature vectors. The enhanced features from different domains are embedded in a common high-level semantic space by a semantic embedding module. Based on the high-level semantic features, three types of losses (i.e., classification loss, improved bidirectional triplet loss, adversarial loss) are proposed to acquire more discriminative cross-domain feature representations. Detailed configuration of these modules can be found in the following sections.

3.2 Mixed high-order attention module

Attention acts as a tool to support more available resources to be allocated to the most important information region of an input. In most FG-SBIR models, attention is adopted to reweight the feature maps to highlight specific discriminative local regions. However, most previous attention models for FG-SBIR are coarse and cannot effectively capture the high-order relationships among parts. In this study, the mixed HOA module is adopted to establish the higher-order relationships among parts and to enhance the richness of attention. For take full advantage of the rich features contained in the middle convolutional layer, we place it between P1 and P2. The mixed HOA module consists of four HOA modules with different orders (i.e. $\{R = 1, 2, 3, 4\}$) which can model and use the diverse high-order information.

We denote the output of mid-level feature space P1 as $F^{conv2} \in R^{H \times W \times C}$. F^{conv2} is first processed by a set of 1×1 convolution layers with weights W_s^r to generate a set of

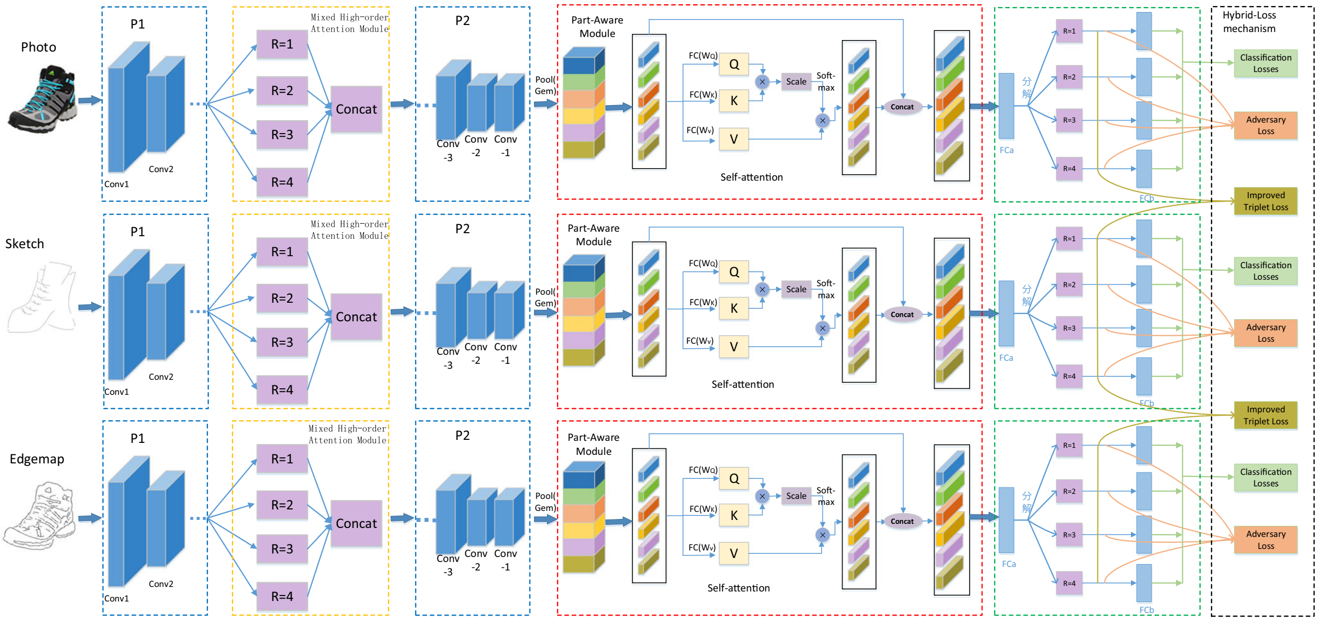


Fig. 4 Architecture of the proposed model

feature maps Z_s^r with channel C^z , which can be formulated as Eq. 3:

$$Z_s^r = W_s^r * F^{conv2} \tag{3}$$

where $r = 1, \cdots, R$ contains all the orders less than or equal to R, $s = 1, \cdots, r$,and $*$ denotes the convolution operator.

Then, a set of element-wise product operations $\odot$ is performed on feature maps $\{Z_s^r\}_{s=1,\cdots,r}$ to obtain Z^r:

$$Z^r = Z_1^r \odot \cdots \odot Z_r^r, r = 1, \cdots, R \tag{4}$$

An r-th order feature map Z^r is processed by non-linear activation function to improve the representation ability of high-order attention. Then a new set of 1×1 convolution layers with weights W_α^r are applied to produce R attention map A^r:

$$A^r = W_\alpha^r * (\sigma(Z^r), r = 1, \cdots, R \tag{5}$$

where σ is ReLU function. Finally $\{A^r\}_{r=1,\cdots,R}$ are combined by sum operation and apply a sigmoid function to obtain the final high-order attention map A:

$$A = sigmoid(\sum_{r=1}^{R} A^r) \tag{6}$$

To overcome the loss of useful information in the feature map F^{conv2} by the final high-order attention map A due to corruption by noise, we use a shortcut connection structure, which combines the input feature F^{conv2} and the output of the HOA module with an element–wise sum. The final output of HOA module is computed as:

$$F^a = A \odot F^{conv2} + F^{conv2} \tag{7}$$

The fourth-order HOA module is shown in Fig. 5.

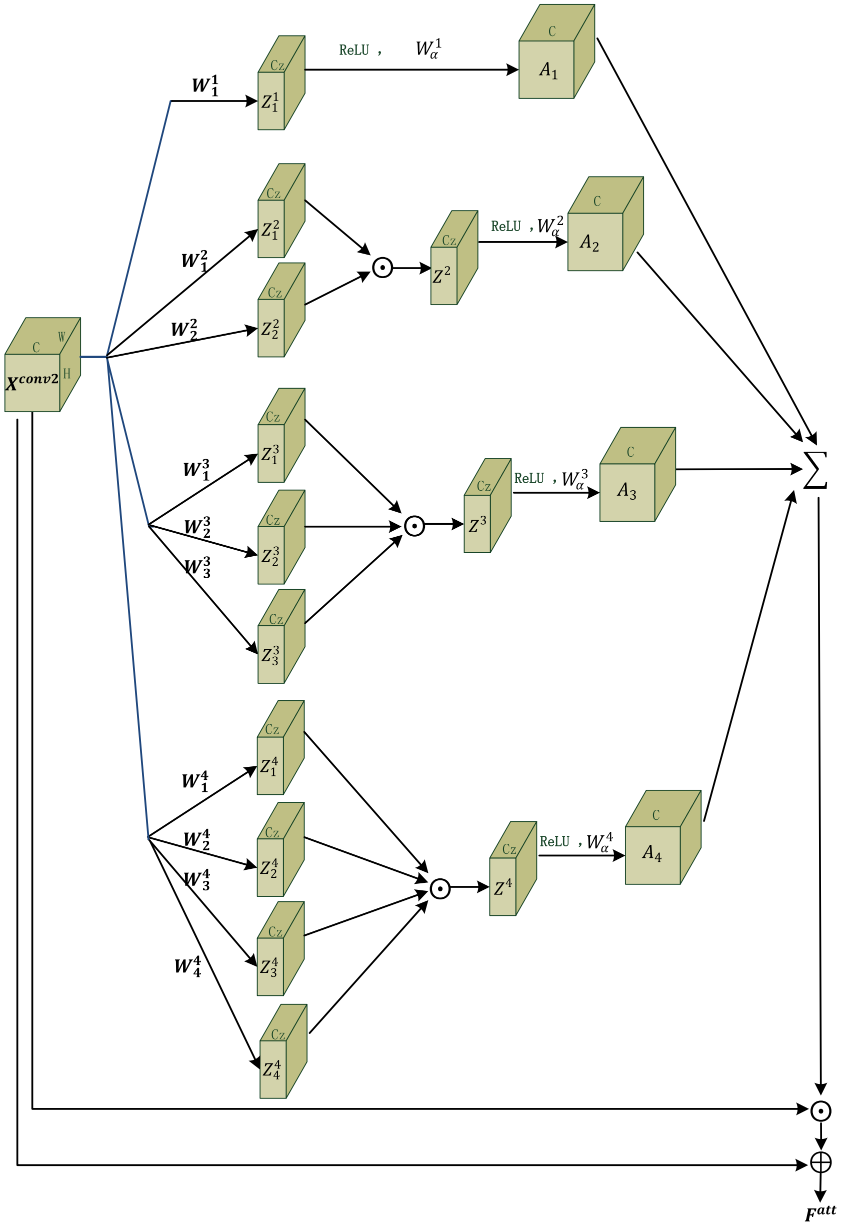
3.3 Part-aware module

Existing FG-SBIR models usually capture a global feature that contains information (e.g., structure, line, position of inflection point), making it easy to ignore small visual clues, such as a bowknot on high heels. However, these details can be useful for inputs with small inter-class variations. Part-level features offer fine-grained information and have been verified as beneficial for image retrieval in very recent literature [47]. Therefore, we design an enhanced part-aware module to obtain more discriminative cross-domain feature representations.

The enhanced part-aware module is shown in Fig. 4. Tensor F^h is the output of the high-level feature space P2. We replace the top original max-pooling layer with a learnable generalized-mean (GeM) pooling layer [48]. We first vertically spilt it into $M = 6$ part to obtain part features $F^p = [F_1^p, F_2^p, \cdots, F_M^p]$, and then flatten it to obtain M local part-level feature vectors $f^p = [f_1^p, f_2^p, \cdots, f_M^p]$. Then we apply a self-attention mechanism [49] on these M part-level feature vectors to learn the relationship among different parts and enhance part specific information contained in each part vectors. This can be formulated as:

$$\begin{aligned} Q &= {W_Q}^{\mathrm{T}} f^p \\ K &= {W_K}^{\mathrm{T}} f^p \\ V &= {W_V}^{\mathrm{T}} f^p \\ f^s &= softmax(\frac{Q \times K^{\mathrm{T}}}{s})V \end{aligned} \tag{8}$$

Fig. 5 Illustration of $R = 4$ HOA module



where W_Q, W_K, W_V are weight matrix of the three FC layer respectively, s is scale factor.

Later, we repeat each part vector into the same spatial shape as F_i^p to obtain part-aware features $F^a = [F_1^a, F_2^a, \cdots, F_M^a]$. Finally, we fuse these part-aware features and original part-level features F^p by channel-wise concatenation to generate the enhanced part-aware features $F^e = [F_1^e, F_2^e, \cdots, F_M^e]$, which serve as the input to the next layer of the network.

3.4 Hybrid-loss mechanism

We hope that the learned feature vector contains not only the visual information that can express the shape and contour changes but also the semantic information that can express the sub-category of the image. We introduce a hybrid-loss mechanism which contains improved bidirectional triplet loss, classification loss, and adversarial loss to optimize the network, while achieving a more discriminative embedding. [50] proved that if we use classification loss and triplet loss to optimize in the same feature space, their goals may be inconsistent. Therefore, two fully connected layers(FCa and FCb) are used as a semantic embedding part to develop a features FC layer and a classifier FC layer, and add a BN layer [51] between them. The feature before the BN layers is denoted as f^t, the feature after the classifier FC layers is denoted as f^c. In the training phase, f^t and f^c are used to calculate triplet loss and classification loss, respectively. As for prevent the order collapse problem, the adversarial loss is introduced to the mixed high-order attention module.

Classification Loss In order to reinforce semantic consistency of local parts, we establish a class label mapping relationship among different domains. Give sketch sample $\{s_i\}_{i=1}^{N}$ and photo sample $\{p_j\}_{j=1}^{N}$, we use the method of [52] to get edgemap sample $\{e_j\}_{j=1}^{N}$. We assign the indexes for the sketches as the class labels. For example, the class label for the sketch s_i is $y_i = i$. The positive photos (or edge maps) have the same class labels as their corresponding sketches. In this paper, we employ classification loss to emphasize the semantic consistency of each part. For each branch, a softmax cross-entropy loss is exploited as part classifier as follows:

$$L_{cls}(f^c) = -\sum_{k=1}^{K} y_k \log(\frac{\exp({f^c}_k)}{\sum_{j=1}^{K}\exp({f^c}_j)}) \tag{9}$$

where K represents the number of categories, the ${f^c}_j$ is the j-element of the prediction score f^c.

The final classification loss is the sum of part classifier is defined as:

$$L_{cls} = \sum_{i=1}^{M}\sum_{j=1}^{R} L_{cls}(f^c(ij)) \tag{10}$$

where $f^c(ij)$ denotes the prediction score of ij part. In this way, the positive samples and sketches are aligned in the high-level semantic space.

Improved Bidirectional Triplet Loss The conventional triplet loss approach only that the distance between matched pairs is smaller than that of non-matched pairs by at least a predefined margin, but it does not specify how close the positive pairs should be. For instance, when $D(f^t(s), f^t(p^+)) = 0.4$, $D(f^t(s), f^t(p^-)) = 0.6$ and $D(f^t(s), f^t(p^+)) = 1.2$, $D(f^t(s), f^t(p^-)) = 1.4$, the triplet loss are both 0.1, although the ranking results are the same. The latter case may result in a relatively large average intra-class distance. In addition, the commonly used triplet loss only emphasizes the unidirectional constraint for two modalities, and ignores the constraint relationships among triplets (p, s^+, s^-).

To overcome these limitations, an improved bidirectional triplet loss function is proposed to enhances the inter-class separability and the intra-class compactness. It is formulated as follows:

$$\begin{aligned} L_{tri}(s,p) =& [D(f^t(s), f^t(p^+)) - D(f^t(s), f^t(p^-)) + \varDelta_1]_+ \\ &+ [D(f^t(p), f^t(s^+)) - D(f^t(p), f^t(s^-)) + \varDelta_1]_+ \\ &+ \beta[D(f^t(s), f^t(p^+)) - \varDelta_2]_+ \end{aligned} \tag{11}$$

where β is the hyper-parameter, the third term ensures that the intra-class distance $D(f^t(s), f^t(p^+))$ is less than a second margin $\varDelta_2$, and that $\varDelta_2$ is much smaller than $\varDelta_1$. The formulation effectively pulls the positive pairs closer, and pushes the negative pairs farther and at the same time increases intra-class compactness.

Additionally, considering the one-to-one correlation between a photo and its corresponding edge map, sketches and edge maps should satisfy the triplet constraint. The final improved bidirectional triplet loss is formulated as follows:

$$\begin{aligned} L_{tri} &= L_{tri}(s,p) + L_{tri}(s,e) \\ &= [D(f^t(s), f^t(p^+)) - D(f^t(s), f^t(p^-)) + \varDelta_1]_+ \\ &\quad + [D(f^t(p), f^t(s^+)) - D(f^t(p), f^t(s^-)) + \varDelta_1]_+ \\ &\quad + \beta[D(f^t(s), f^t(p^+)) - \varDelta_2]_+ \\ &\quad + [D(f^t(s), f^t(e^+)) - D(f^t(s), f^t(e^-)) + \varDelta_1]_+ \\ &\quad + [D(f^t(e), f^t(s^+)) - D(f^t(e), f^t(s^-)) + \varDelta_1]_+ \\ &\quad + \beta[D(f^t(s), f^t(e^+)) - \varDelta_2]_+ \end{aligned} \tag{12}$$

Adversarial Loss The biased learning behavior of the deep model may cause a high-order HOA module to collapse into a relatively low-order module, so that the mixed high-order attention module cannot capture the expected diverse higher-order attention information and cannot achieve optimal performance. The adversary constraint [30] adjusts the order of HOA to be different, which can be formulated as:

$$\max_{HOA|_{R=1}^{R=K}} \min_{F}(L_{adv}) = \max_{HOA|_{R=1}^{R=K}} \min_{F}(\sum_{j,j^{`},j\neq j^{`}}^{k} D(F(f_j), F(f_{j^{`}}))) \tag{13}$$

where $HOA|_{R=1}^{R=K}$ denotes k HOA modules from first-order to k-th order. $F(\cdot)$ is an encoding function composed of two fully-connected layers. f_j is the output of the HOA module with $R = j$. By playing the max–min game as in [27], the problem of order collapse can be suppressed.

In summary, the overall objective function of our network can be formulated as:

$$\min(L_{cls} + \lambda_1 L_{tri}) + \lambda_2 \max_{HOA|_{R=1}^{R=K}} \min_{F}(L_{adv}) \tag{14}$$

where λ_1, λ_2 represent the hyper-parameters.

4 Experiments

4.1 Dataset and evaluation protocol

Datasets We choose several popular FG-SBIR benchmarks: QMUL-Shoe, QMUL-Chair, and QMUL-Handbag, QMUL-Shoes-V2 and Sketchy as our experimental datasets.

Table 1 The training/testing splits of FG-SBIR datasets

Datasets	Training Photos/Sketches	Testing Photos/Sketches
QMUL-Shoe	304/304	115/115
QMUL-Chair	200/200	97/97
QMUL-Handbag	400/400	168/168
QMUL-Shoes-V2	1800/5982	200/666
Sketchy	11250/68113	1250/6312

The first three datasets contain only one category of sketch-photo pairs, that is, shoe/chair/handbag category respectively. QMUL-Shoes-V2 dataset is an extension of the QMUL-Shoe dataset that includes more sketches and photos. Sketchy is a large-scale dataset that contains sketch–photo data span 125 categories, each photo corresponds to 5-20 sketches. In these datasets, our training triplets are automatically generated: for each anchor sketch, the true matching natural photo form the positive pair, while the negative pair is randomly sampled from all other images. We split the samples in each dataset for training and testing according to Table 1.

Evaluation Protocols For FG-SBIR task, the evaluation metrics we used is the same recall *@K* as in [5]. Give one query sketch, if the correct photo is ranked in the top K, the recall *@K* is one, otherwise it is zero. *acc@K* is the average of all query results.

4.2 Implementation details

Our model is performed on Pytorch using a single NVIDIA 1080Ti GPU. During training, the size of input photos is adjusted to 225 × 225. Data augmentation methods used were: random cropping, random flipping, and random erasing [53], where the probability of dropout is set to 0.4. We use SGD as optimizer with a batch size of 30, the initial learning rate is set as 0.005 for AlexNet base layers and 0.05 for the others in the first 300 epochs, and further decreased to 10^{-4} for another 500 epochs. We set hyper-parameters $\lambda_1 = 1, \lambda_2 = 1, \beta = 0.1, \Delta_1 = 0.3, \Delta_2 = 0.05$.

4.3 Comparative results

Baselines we choose four baseline models for comparison. (1)**TripletSN** [5] used the Sketch-a-Net [19] to form a three branches Siamese network to learn features of both edge maps and sketches, and used the traditional first-order Euclidean triplet loss to optimize. (2)**DSSA** [4] made the following improvements on the TripletSN to improve the performance: added an attention mechanism, built a coarse-fine fusion block, and used higher-order HOELF loss to replace conventional first-order energy function. (3)**EdgeMAC** [22] turned images into edge maps first, and then captured a global image descriptor by a fully convolutional network for training.(4)**DeepTCNet** [32] employed DenseNet-169 as the feature extractor in the Siamese network and adopted the triplet loss and classification loss to optimize. (5)**GNTriplet** [54] is a triplet network based on GoogLeNet and trained with Triplet and Classification loss.
Results Table 2 lists the results of comparing the performance of our method to those of the baseline networks on the three benchmark datasets. We highlight the best results in bold, and the bolds in the table below have the same effect. As can be seen from the table: (1) overall, the performance of our network over the three datasets is better than all baseline models. The improvement is especially obvious on the handbag dataset—an approximately 13%

Table 2 Comparative results against baselines on QMUL-Shoe-Chair-Handbag dataset

Models	QMUL-Shoe (%)		QMUL-Chair (%)		QMUL-Handbag (%)	
	acc@1	*acc@10*	*acc@1*	*acc@10*	*acc@1*	*acc@10*
TripletSN	52.17	92.17	72.16	98.96	39.88	82.14
DSSA	61.74	94.78	81.44	95.88	49.40	82.74
EdgeMAC	54.78	92.17	85.57	97.94	51.19	85.71
DeepTCNet	63.48	95.65	**95.88**	**100**	-	-
Ours	**68.70**	**98.26**	**95.88**	**100**	**64.88**	**89.88**

Table 3 Comparative results on QMUL-Shoes-V2 and Sketchy dataset

Models	QMUL-Shoes-V2	Sketchy
	acc@1 (%)	*acc@1(%)*
TripletSN	30.93	21.63
DSSA	33.63	-
DeepTCNet	**40.02**	**40.81**
GNTriplet	21.62	37.10
Ours	36.79	38.98

increase in top-1 accuracy against the second-best model. Each sketch may have several visually similar photos, the excellent performances of our model at *acc@1* show that the model can identify fine-grained differences between candidate photos. (2) Compared with the sketches of shoes and handbags, the improvement in the identification of chair sketches is less prominent. The reason for this is clear. With its enhanced part-aware module, our model can focus on discriminative local parts; chairs contain relatively few local visual cues compared to shoes and handbags. (3) Both DeepTCNet and our model adopt triplet loss and classification loss to optimize together, and their performances are better than the other three baselines. This demonstrates that the combination of the two losses can lead to better learning of feature representations in FG-SBIR.

We additionally evaluate them on QMUL-Shoes-V2 and Sketchy datasets, results in Table 3. We can know that: (1) The results of our model are superior to edgemap-based methods such as TripletSN and DSSA. The input images of DSSA and TripletSN are edgemaps rather than RGB photos.

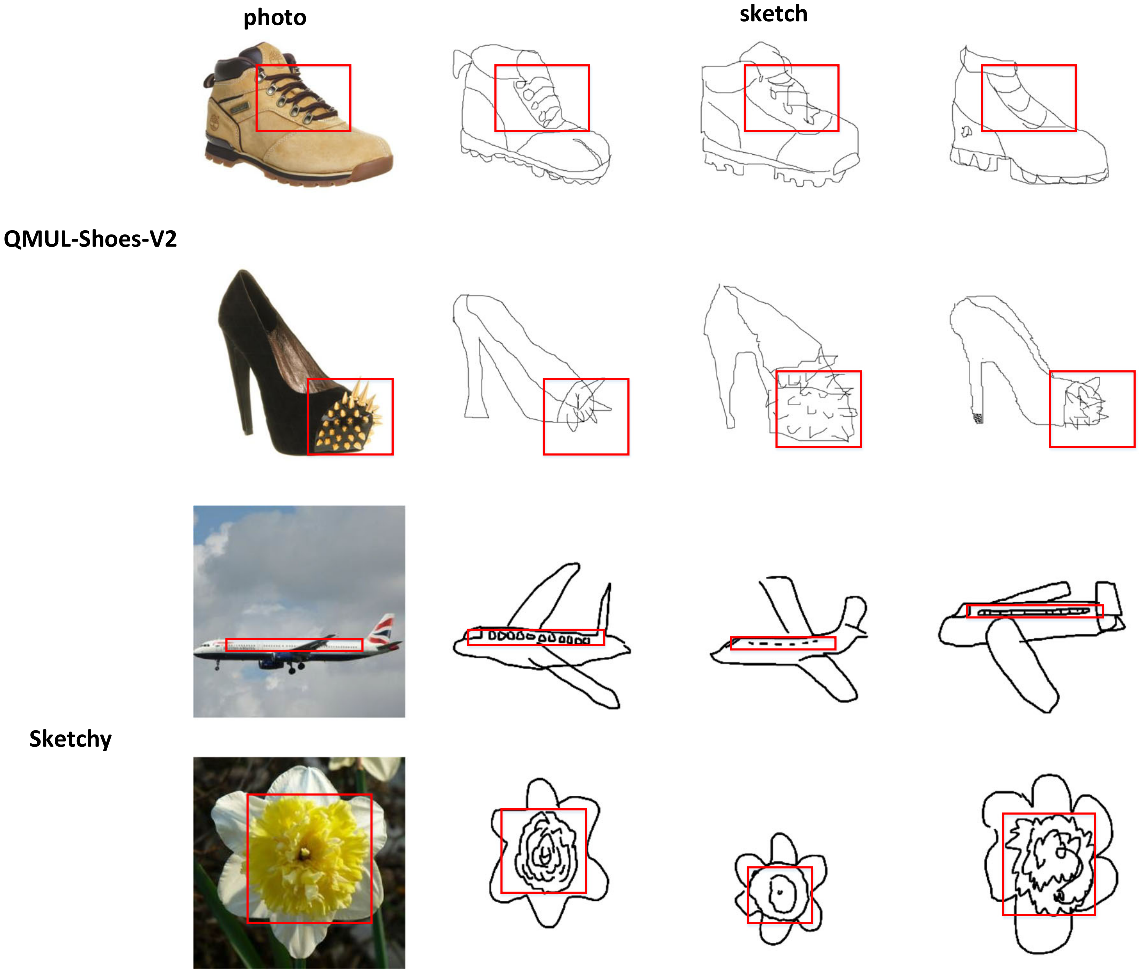


Fig. 6 Examples of photo-sketch pairs in QMUL-Shoes-V2 and Sketchy datasets. Detail parts are circled by red rectangle. We can see that sketches corresponding to the same photo have different descriptions of details

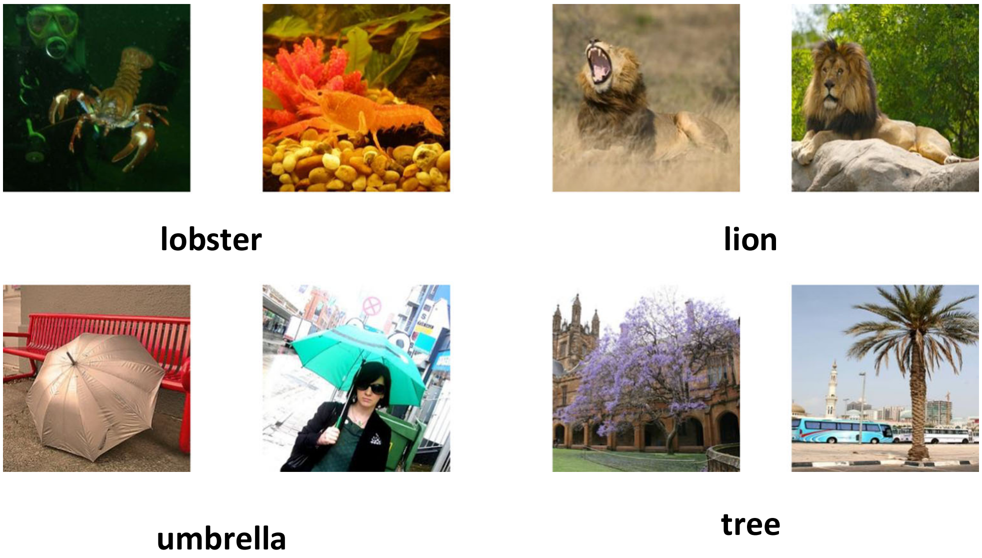


Fig. 7 Examples of photos in Sketchy dataset

To some extent, extracting edge images can effectively reduce the gap between photo domain and sketch domain and improve the retrieval effect. However, the conversion from RGB photos to edge maps may lose some important details for fine-grained retrieval, which makes it difficult to further improve the retrieval results. We added the photo branch to expand the network into three-way, further improving the performance of fine-grained retrieval. (2) Compared with GNTriplet, the improvement of the results indicates that the mixed high-order attention module and the part-aware module added in our network have achieved certain results. (3) The performance of our model is slightly lower than DeepTCNet, which takes into account not only the constraints of feature vector in European space but also in angular space.

From Tables 2 and 3, we find that the performance of our network on QMUL-Shoes-V2 and Sketchy datasets is weaker than that on QMUL-Shoe-Chair-Handbag datasets. This is because: (1) there are multiple sketches corresponding to one photo in the Shoes-V2 and Sketchy datasets. For the same detail part in the photo, the descriptions of multiple sketches are different due to different painting habits and skills of painters, which increases the difficulty of our retrieval. As shown in the Fig. 6. (2) The photos in Sketchy dataset contain many backgrounds (as shown in Fig. 7.), and edge maps extracted by edge detection method also has some irrelevant backgrounds, which will bring some interference and affect the retrieval performance. Overall, the results verify the fine-grained retrieval capability of our model.

4.4 Ablation study

Contributions of key Components We compare our full model with three variants to investigate the contribution of each key component in the former. The experimental results

Table 4 Contributions of the key components

Datasets	Models	*acc@1*	*acc@10*
	Base	55.65	93.04
	w/o MHA	66.09	97.39
	w/o PAM	60.00	95.65
	Full model	**68.70**	**98.26**
QMUL-Chair	Base	85.57	98.97
	w/o MHA	92.78	100
	w/o PAM	88.66	100
	Full model	**95.88**	**100**
QMUL-Handbag	Base	52.98	83.33
	w/o MHA	61.31	89.88
	w/o PAM	57.74	89.29
	Full model	**64.88**	**89.88**

Table 5 Results of Improved bidirectional triplet loss and convolutional triplet loss

(%)	QMUL-Shoe		QMUL-Chair		QMUL-Handbag	
	acc@1	*acc@10*	*acc@1*	*acc@10*	*acc@1*	*acc@10*
Full model with convolutional triplet loss	66.09	97.39	93.81	100	62.50	89.29
Full model with improved triplet loss	**68.70**	**98.26**	**95.88**	**100**	**64.88**	**89.88**

are presented in Table 4. Base refers to the base network, which is optimized by triplet loss and classification loss. "w/o MHA" and "w/o PAM" denote the model without a mixed high-order attention module and without a part-aware module, respectively. Table 4 shows clearly that: (1) The full model outperforms the other three variants on all three datasets, which indicates that the mixed high-order attention module and the part-aware module are complementary to each other. (2) The mixed high-order attention module and the part-aware module both improve the results of base model, demonstrating the contribution of the two components to the overall performance. (3) It is clear that the part-aware module plays a more important role than the mixed high-order attention module for a retrieval task.

Effectiveness of Improved bidirectional triplet loss To further confirm the usefulness of the improved bidirectional triplet loss, we compared it with the traditional triplet loss. Table 5 shows that using the improved bidirectional triplet loss function to train the model, the performance is improved by about 2% over the same model using the conventional triplet loss function.

Further analysis In order to support the theory that the features of the middle convolutional layer play an important role in fine-grained retrieval, we compare the results of the mixed high-order attention module, which placed after different convolutional layers. "MHA-middle" and "MHA-high" denote that the mixed high-order attention module is placed after the middle feature space and the high feature space, respectively. From Table 6, we can see that the accuracy of "MHA-middle" has increased by about 3% compared to the "MHA-high" model. It suggests that for the FG-SBIR task, features contained in the middle convolution layer are indeed important and placing the mixed high-order attention module after the second convolutional layer is the correct choice.

Table 6 Results of placing the mixed high-order attention module at different convolution layer

QMUL-Shoe	*acc@1*	*acc@10*
MHA-middle	**68.70**	**98.26**
MHA-high	66.09	98.26

We plot recall @ K for $K = 1$ to 10 on QMUL-Shoe-Chair-Handbag datasets. The data in Fig. 8 are the average results of 10 tests. It can be found that when the value of K is relatively small, our network can still achieve good results, which suggests that given a query sketch, our model can find a matched image as quickly as possible. This also verifies the great capability of our network on extracting representative and discriminative features.

4.5 Qualitative visualization

For an intuitive understanding of our retrieval results, we present some qualitative examples in Fig. 9. We highlight the correct retrieval results with red rectangles. It can be seen that our model has an obvious advantage in finding fine-grained similarities between sketches and photos. For example, on the first shoe example, when given a shoe with a shoelace as the query sketch, our model can find similar images with a shoelace detail and bring the correct shoe to Rank 1. These qualitative results demonstrate that our

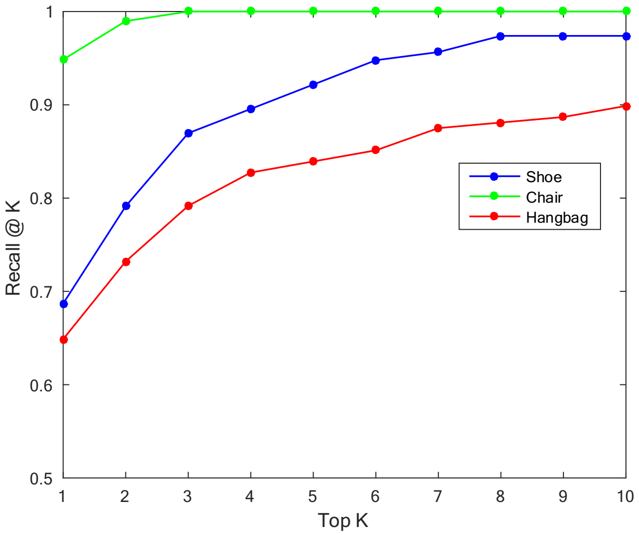


Fig. 8 Evaluation on QMUL-Shoe-Chair-Handbag datasets at Top K. We measure acc@1 whether or not the network can retrieve target image within Top K nearest neighbors

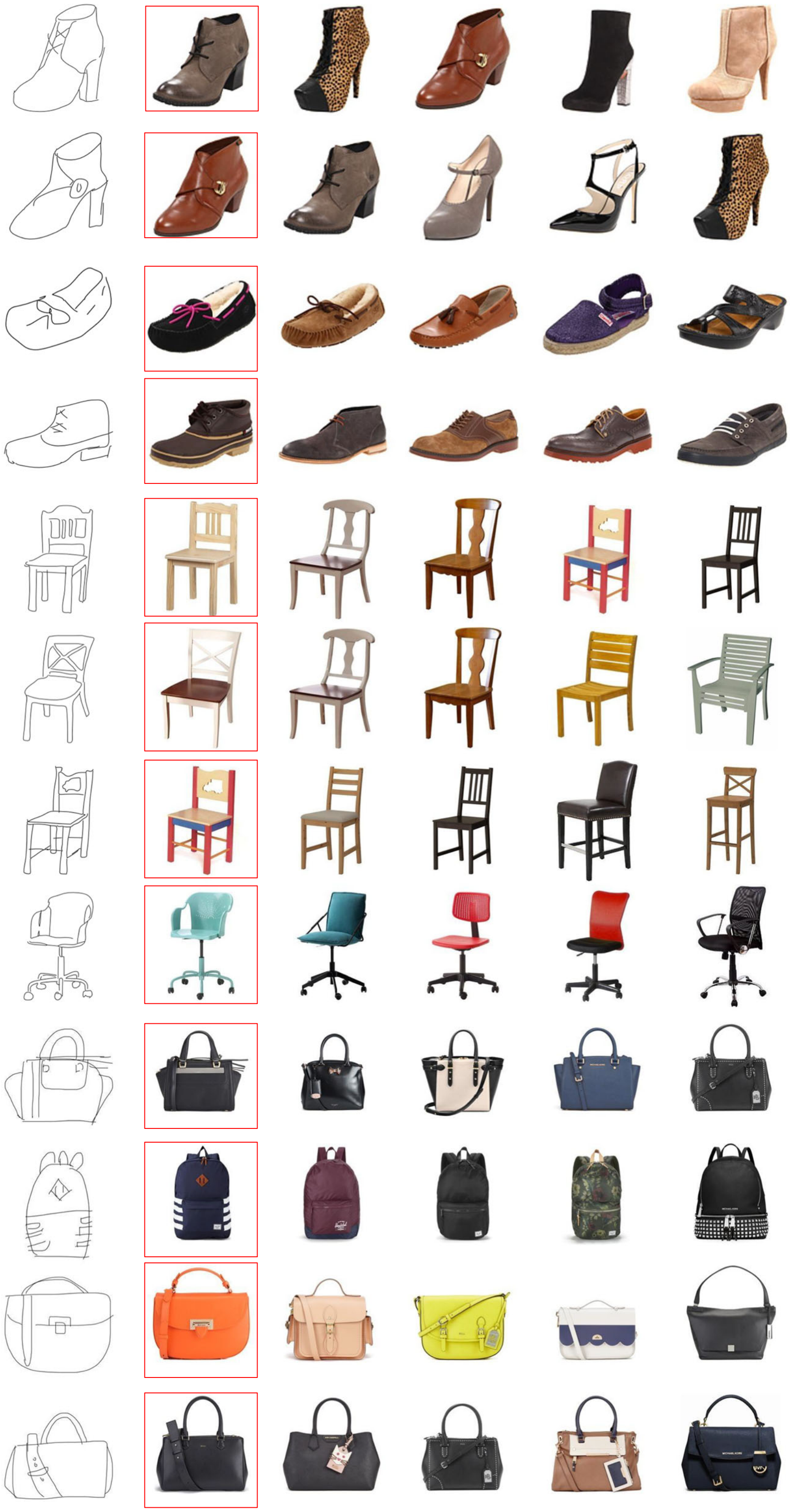


Fig. 9 The top 5 retrieval results of our method. For each example, the first column is query sketch, and the correct retrieval result is highlighted in red rectangle

model can learn more fine-grained features and has better discriminability for the FG-SBIR task.

5 Conclusion

FG-SBIR is a particularly challenging task due to the “one-to-many” mapping relationship between free-hand sketches and natural photos, we propose a novel three-way enhanced part-aware network to address the problem of sketch-based image retrieval, which is of great practical significance in today’s world populated by smart touch-screen devices. The main contribution is the introduction of a mixed high-order module and part-aware module in each branch network, which can capture more useful fine-grained cues and enhance the semantic consistency of local regions. The architecture was optimized by a hybrid-loss and learned more discriminative features. Experimental results on three baseline datasets validated the effectiveness of our model.

However, our model is only able to study sketches from a static pixel space and does not make full use of the timing of sketches. In future research, we will focus our attention on dynamic stroke-coordinate spaces to design a sketch-specific data augmentation approach and model.

Acknowledgements This study was supported by the National Natural Science Foundation of China (No. 61772032) and the National Key R&D Project (SQ2018YFC080102).

References

1. Yangbo FGQYX (2007) Review on technology of text-based image retrieval. Sci Mosaic 3
2. Bhave AM, Wanjari M, Sawarkar G (2014) Iretrieval: Image retrieval based on color feature and texture feature. Int J Adv Res Comput Sci 5(6)
3. Smeulders AW, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22(12):1349–1380
4. Song J, Yu Q, Song YZ, Xiang T, Hospedales TM (2017) Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In: Proceedings of the IEEE international conference on computer vision, pp 5551–5560
5. Yu Q, Liu F, Song YZ, Xiang T, Hospedales T, Loy CC (2016) Sketch me that shoe. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 799–807
6. Liu L, Shen F, Shen Y, Liu X, Shao L (2017) Deep sketch hashing: Fast free-hand sketch-based image retrieval. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2862–2871
7. Sangkloy P, Burnell N, Ham C, Hays J (2016) The sketchy database: learning to retrieve badly drawn bunnies. ACM Trans Graph (TOG) 35(4):1–12
8. Qi Y, Song YZ, Zhang H, Liu J (2016) Sketch-based image retrieval via Siamese convolutional neural network. In: 2016 IEEE International conference on image processing (ICIP), pp 2460–2464
9. Li K, Pang K, Song YZ, Hospedales T, Zhang H, Hu Y (2016) Fine-grained sketch-based image retrieval: The role of part-aware attributes. In: 2016 IEEE Winter conference on app lications of computer vision (WACV), pp 1–9
10. Saavedra JM, Bustos B (2014) Sketch-based image retrieval using keyshapes. Multimed Tools Appl 73(3):2033–2062
11. Eitz M, Hildebrand K, Boubekeur T, Alexa M (2010) Sketch-based image retrieval: Benchmark and bag-of-features descriptors. IEEE Trans Vis Comput Graph 17(11):1624–1636
12. Hu R, Collomosse J (2013) A performance evaluation of gradient field hog descriptor for sketch based image retrieval. Comput Vis Image Underst 117(7):790–806
13. Cao Y, Wang C, Zhang L, Zhang L (2011) Edgel index for large-scale sketch-based image search. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, pp 761–768
14. Eitz M, Hildebrand K, Boubekeur T, Alexa M (2010) An evaluation of descriptors for large-scale image retrieval from sketched feature lines. Comput Graph 34(5):482–498
15. Zhou R, Chen L, Zhang L (2012) Sketch-based image retrieval on a large scale database. In: Proceedings of the 20th ACM international conference on Multimedia, pp 973–976
16. Lei J, Song Y, Peng B, Ma Z, Shao L, Song YZ (2019) Semi-heterogeneous three-way joint embedding network for sketch-based image retrieval. IEEE Trans Circ Syst Video Technol 30(9):3226–3237
17. Parui S, Mittal A (2014) Similarity-invariant sketch-based image retrieval in large databases. In: European conference on computer vision, springer, pp 398–414
18. Li K, Pang K, Song YZ, Hospedales T, Xiang T, Zhang H (2017) Synergistic instance-level subspace alignment for fine-grained sketch-based image retrieval. IEEE Trans Image Process 26(12):5908–5921
19. Yu Q, Yang Y, Liu F, Song YZ, Xiang T, Hospedales TM (2017) Sketch-a-net: a deep neural network that beats humans. Int J Comput Vis 122(3):411–425
20. Zhang H, Zhang C, Wu M (2017) Sketch-based cross-domain image retrieval via heterogeneous network. In: 2017 IEEE Visual communications and image processing (VCIP), pp 1–4
21. Bui T, Ribeiro L, Ponti M, Collomosse J (2018) Sketching out the details:Sketch-based image retrieval using convolutional neural networks with multi-stage regression. Comput Graph 71:77–87
22. Radenovic F, Tolias G, Chum O (2018) Deep shape matching. In: Proceedings of the European conference on computer vision (eccv), pp 751–767
23. Song J, Pang K, Song YZ, Xiang T, Hospedales T (2018) Learning to sketch with shortcut cycle consistency. In: 2018 IEEE/CVF Conference on computer vision and pattern recognition, pp 801–810
24. Pang K, Song YZ, Xiang T, Hospedales TM (2017) Cross-domain generative learning for fine-grained sketch-based image retrieval. In: BMVC, pp 1–12
25. Ha D, Eck D (2017) A neural representation of sketch drawings. arXiv:170403477
26. Chen W, Hays J (2018) Sketchygan: Towards diverse and realistic sketch to image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 9416–9425
27. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial networks. arXiv:14062661
28. Yu Q, Chang X, Song YZ, Xiang T, Hospedales TM (2017) The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching. arXiv:171108106

29. Hermans A, Beyer L, Leibe B (2017) In defense of the triplet loss for person re-identification. arXiv:170307737
30. Chen B, Deng W, Hu J (2019) Mixed high-order attention network for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 371–381
31. Li Y, Hospedales T, Song YZ, Gong S (2014) Fine-grained sketch-based image retrieval by matching deformable part models. In: The british machine vision conference (BMVC)
32. Lin H, Fu Y, Lu P, Gong S, Xue X, Jiang YG (2019) TC-Net for iSBIR: Triplet classification network for instance-level sketch based image retrieval. In: Proceedings of the 27th ACM International Conference on Multimedia, pp 1676–1684
33. Pang K, Li D, Song J, Song YZ, Xiang T, Hospedales TM (2018) Deep factorised inverse-sketching. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 36–52
34. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:181004805
35. Ribeiro LSF, Bui T, Collomosse J, Ponti M (2020) Sketchformer:transformer-based representation for sketched structure. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14153–14162
36. Lin H, Fu Y, Xue X, Jiang YG (2020) Sketch-bert: Learning sketch bidirectional encoder representation from transformers by self-supervised learning of sketch gestalt. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 6758–6767
37. Pang K, Li K, Yang Y, Zhang H, Hospedales T, Xiang T, Song YZ (2019) Generalising fine-grained sketch-based image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 677–686
38. Bhunia AK, Yang Y, Hospedales T, Xiang T, Song YZ (2020) Sketch less for more:On-the-fly fine-grained sketch-based image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 9779–9788
39. Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, Wang X, Tang X (2017) Residual attention network for image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3156–3164
40. Li W, Zhu X, Gong S (2018) Harmonious attention network for person re-identification. In: 2018 IEEE/CVF Conference on computer vision and pattern recognition, pp 2285–2294
41. Li D, Chen X, Zhang Z, body HuangK (2017) Latent parts for person re-identification Learning deep context-aware features over. In: Proceedings of the IEEE conference on computer vision and pattern recognition pp 384–393
42. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
43. Yu R, Dou Z, Bai S, Zhang Z, Xu Y, Bai X (2018) Hard-aware point-to-set deep metric for person re-identification. In: Proceedings of the European conference on computer vision (ECCV), pp 188–204
44. Chen W, Chen X, Zhang J, Huang K (2017) Beyond triplet loss: a deep quadruplet network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 403–412
45. Shi H, Yang Y, Zhu X, Liao S, Lei Z, Zheng W, Li SZ (2016) Embedding deep metric for personre-identification: A study against large variations. In: European conference on computer vision. Springer pp 732–748
46. Chopra S, Hadsell R, LeCun Y (2005) Learning a similarity metric discriminatively, with app lication to face verification. In: 2005 IEEE Computer society conference on computer vision and pattern recognition(CVPR), vol 1, pp 539–546
47. Sun Y, Zheng L, Yang Y, Tian Q, Wang S (2018) Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European conference on computer vision (ECCV), pp 480–496
48. Radenović F, Tolias G, Chum O (2018) Fine-tuning CNN image Image retrieval with no human annotation. IEEE Trans Pattern Anal Mach Intell 41(7):1655–1668
49. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. arXiv:170603762
50. Luo H, Jiang W, Gu Y, Liu F, Liao X, Lai S, Gu J (2019) A strong baseline and batch normalization neck for deep person re-identification. IEEE Trans Multimed 22(10):2597–2609
51. Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning, PMLR 37:448–456
52. Zitnick CL, Dollár P (2014) Edge boxes: Locating object proposals from edges. In: European conference on computer vision, Springer, pp 391–405
53. Zhong Z, Zheng L, Kang G, Li S, Yang Y. (2020) Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 34 pp 13001–13008
54. Sangkloy P, Burnell N, Ham C, Hays J. (2016) The sketchy database:Learning to retrieve baddly drawn bunnies. ACM Transactions on Graphics (proceedings of SIGGRAPH)

Xingying Wang is currently pursuing the Ph.D. degree with the School of Electronics and Information Engineering, Anhui University, Hefei, China. Her research interests include computer vision, machine learning, and pattern recognition.

Jun Tang is currently a professor with the School of Electronics and Information Engineering, Anhui University, Hefei, China. His research interests include computer vision, machine learning, and pattern recognition.

Shoubiao Tan received the Ph.D. degree from the University of Science and Technology of China, in 2004. He is currently a Professor with Anhui University, China. His current research interests include computer vision and pattern recognition.