Deep Spatial-Semantic Attention for Fine-Grained Sketch-Based Image Retrieval

Deep Spatial-Semantic Attention for Fine-Grained Sketch-Based Image Retrieval

Jifei Song* Qian Yu* Yi-Zhe Song Tao Xiang Timothy M. Hospedales
Queen Mary University of London University of Edinburgh
{j.song, q.yu, yizhe.song, t.xiang}@qmul.ac.uk, t.hospedales@ed.ac.uk

Abstract

Human sketches are unique in being able to capture both the spatial topology of a visual object, as well as its subtle appearance details. Fine-grained sketch-based image retrieval (FG-SBIR) importantly leverages on such fine-grained characteristics of sketches to conduct instance-level retrieval of photos. Nevertheless, human sketches are often highly abstract and iconic, resulting in severe misalignments with candidate photos which in turn make subtle visual detail matching difficult. Existing FG-SBIR approaches focus only on coarse holistic matching via deep cross-domain representation learning, yet ignore explicitly accounting for fine-grained details and their spatial context. In this paper, a novel deep FG-SBIR model is proposed which differs significantly from the existing models in that: (1) It is spatially aware, achieved by introducing an attention module that is sensitive to the spatial position of visual details; (2) It combines coarse and fine semantic information via a shortcut connection fusion block; and (3) It models feature correlation and is robust to misalignments between the extracted features across the two domains by introducing a novel higher-order learnable energy function (HOLEF) based loss. Extensive experiments show that the proposed deep spatial-semantic attention model significantly outperforms the state-of-the-art.

1. Introduction

With the proliferation of touch-screen devices, a number of sketch-based computer vision problems have attracted increasing attention, including sketch recognition [47, 36, 3, 32], sketch-based image retrieval [46, 24, 10], sketch-based 3D model retrieval [39], and forensic sketch analysis [14, 28]. Among them, using a sketch to retrieve a specific object instance, or fine-grained sketch-based image retrieval (FG-SBIR) [15, 46, 31] is of particular interest due to its potential in commercial applications such as searching online product catalogues for shoes, furniture, and handbags by finger-sketching on a smart-phone screen.

Figure 1. FG-SBIR is challenging due to the misalignment of the domains (left) and subtle local appearance differences between a true match photo and a visually similar incorrect match (right).

FG-SBIR is a very challenging problem and remains unsolved. First, there is a large domain gap between sketch and photo – a sketch captures mainly object shape/contour information and contains no information on colour and very little on texture. Second, FG-SBIR is typically based on free-hand sketches which are drawn based on mental recollection of reference images shown moments before the drawing stage, making free-hand sketches distinctly more abstract than line tracings (human edgemaps). As a result, a sketch and its matched photo could have large discrepancies in shape and spatial misalignment both globally and locally. Finally, as an object instance recognition problem, given a query sketch, there are often many visually similar candidate photos in the gallery; the correct match and wrong matches may only differ subtly in some localised object parts. Some of these challenges are illustrated in Fig. 1.

Existing FG-SBIR models focus primarily on closing the semantic gap between the two domains whilst only partially addressing or completely ignoring the latter two challenges. Specifically, state-of-the-art FG-SBIR models [46, 31] adopt a multi-branch deep convolutional neural networks (CNNs). Each domain has a corresponding branch which consists of multiple convolutional/pooling layers followed by fully connected (FC) layers. The final FC layer is used as input to pairwise verification or triplet ranking losses to align the domains. However, recent efforts [6, 22] on visualising what each layer of a CNN actually learns

* These authors contributed equally to this work

show that higher-layers of the network capture more abstract semantic concepts but not fine-grained detail, motivating fine-grained recognition methods to work with convolutional feature maps instead [17]. After many pooling and FC layers, the spatial fine-grained details is gone and cannot be recovered. Thus existing deep FG-SBIR models are unable to tell apart visually similar photos based on subtle differences.

In this paper, we introduce spatial-semantic attention modelling in deep FG-SBIR. The architecture of the proposed model is shown in Fig. 2. Although it is still essentially a multi-branch CNN, there are a number of crucial differences to existing models. First, we introduce attention modelling in each branch of the CNN so that computation for representation learning is focused on specific discriminative local regions rather than being spread evenly over the whole image. Due to the large misalignment between the sketch and photo domains, directly taking the attended feature map as input to the subsequent layers of the network is too sensitive to misalignment. We thus introduce a shortcut connection architecture [37, 8] to link the input directly to the output of the attention module so that an imprecise attention mask would not derail the deep feature computation completely, resulting in robust attention modelling. Second, we keep both coarse and fine semantic details through another shortcut block to connect the attended feature map with the final FC layer before feeding it to the loss.

Including fine-grained information in the CNN feature output enables discrimination based on subtle details, but has two risks: misalignment in the feature channels between the two branches, and greater feature noise due to each fine-grained feature having less supporting cues. Existing pairwise verification or triplet ranking losses [46, 31] are sensitive to misalignment. Specifically, these losses typically use Euclidean distance based energy function which relies on element-wise distance computation. They thus implicitly assume that the compared feature vectors are *perfectly element-wise aligned*, an assumption that is violated in practice. To overcome these problems, we propose a novel higher-order learnable energy function (HOLEF) based loss. Using this energy function, when comparing a sketch and photo, the outer subtraction between the two feature vectors are computed, exhaustively measuring the element-wise feature difference across the two domains. This allows increased sensitivity without loosing robustness, by accounting for common misalignments, via learning to exploit any systematic co-occurrences of feature activations in both branches, and using correlated activations to provide robustness to noise.

2. Related Work

Fine-grained SBIR Most existing SBIR works [23, 24, 9, 1, 2, 38, 11, 18, 13, 39, 10, 48] focus on category-level sketch-to-photo retrieval. The problem of fine-grained SBIR was first proposed in [15], which employed a deformable part-based model (DPM) representation and graph matching. More recently, the FG-SBIR problem is tackled by deep learning [46, 31] which aims to learn both feature representation and cross-domain matching function jointly. Both models in [46, 31] evaluated two-branch CNNs with pairwise verification loss and three-branch CNNs with triplet ranking loss and concluded that the latter is better. They differ in whether the network is Siamese or heterogeneous. The model in [46] is Siamese as it takes as input extracted edge maps for the photo branch, whilst the model in [31] is heterogeneous without the edge extraction operation. Our network is also a three-branch CNN. But with the introduced attention modelling, multi-scale coarse-fine semantic fusion, and HOLEF loss, our model is much more effective as validated by our experiments (see Sec. 4).

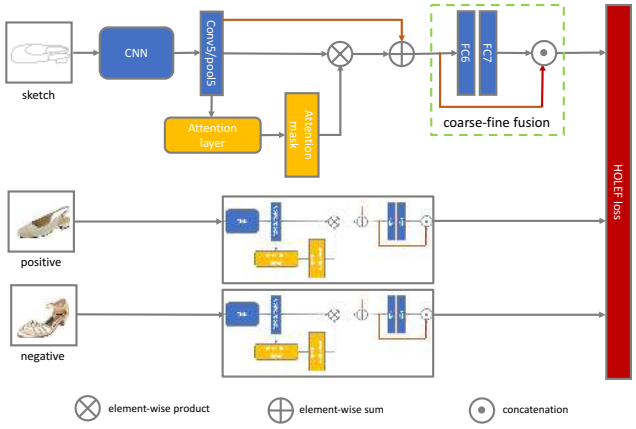

Figure 2. Architecture of the proposed model.

Attention Modelling Visual attention models have been studied extensively in a wide range of vision problems including image caption generation [44, 20], VQA [5, 26], image classification [25, 34, 42] and particularly fine-grained image recognition [34, 42]. Various types of attention models exist. Soft attention is the most commonly used one because it is differentiable thus can be learned end-to-end with the rest of the network. Most soft-attention models learn an attention mask which assigns different weights to different regions of an image. Alternatively, the spatial transformer network [12] generates an affine transformation matrix which locates the discriminative region. Different from soft attention, hard attention models only indicate one region at each time. A hard attention model is not differentiable so it is typically learned using reinforcement learning. Interestingly, there is no prior SBIR (both category-level and instance-level) work that models attention, perhaps because conventional attention models deployed in a cross-domain match problem assume pixel-level alignment; they

thus become ineffective when this assumption is invalid as in the case of SBIR. Our attention model is specifically designed for FG-SBIR in that it is robust against spatial misalignment through the shortcut connection architecture.

Shortcuts and Layer Fusion in Deep Learning The shortcut architecture used in both the attention module and the coarse-fine fusion block in our model serve to fuse multiple layers at different depths. Fusing different CNN layers in the model output has been exploited in many problems such as edge detection (e.g., [30, 43]), pose estimation (e.g., [27]) and scene classification (e.g., [7, 45, 19]. The motivation is typically multi-scale (coarse to fine) fusion rather than attended-unattended feature map fusion, as in our first shortcut block. Various shortcut connection architectures have been successfully deployed in a number of widely used CNNs including GoogLeNet [37] and ResNet [8]. Our shortcut connection architecture is similar to that of the residual block in ResNet [8]. However, instead of making the network deeper, we use it in the attention module to make the attention module output robust against imprecise attention mask caused by cross-domain feature misalignment, as well as in the final CNN output layer to preserve both coarse and fine-grained information in the learned representation.

Higher-order Energy Function Loss functions for verification or ranking typically use an energy function, that measures the (dis)similarity between two feature vectors. For example, triplet loss is widely used in many deep verification [33, 29] or ranking [40, 46, 36, 31] networks. It is adopted here to enforce the ranking between a query sketch and a pair of positive and negative photos. In the vast majority of cases [40, 46, 36, 31] Euclidean distance-based, or other first-order energy functions are used in the loss formulation. They are first order in the sense that only element-wise comparisons are made, making it sensitive to feature misalignment and meaning that no cross-feature correlation can be exploited in the similarity. The proposed HOLEF loss is a triplet loss with a 2nd-order energy function based on a weighted outer subtraction between a pair of input vectors. Compared to first-order alternatives, our energy function is more robust against misalignment between sketch and photo channels, and can accommodate better the more detailed but noisier fine-grained feature map representation. Mahalanobis distance [41, 35] is another example of a higher-order energy function in that it does $O(N^2)$ comparisons for N channels. However it is based on element-wise difference followed by bilinear product so the effect is to learn which dimension pairs are important to match, rather than compensate for misalignment and noise between the input vectors.

The Contributions of this work are as follows: (1) A novel deep FG-SBIR model is proposed. The model learns discriminative feature representation that is spatially attended and includes both coarse and fine details. (2) A new higher-order learnable energy function (HOELF) based loss is used to make the model robust against feature misalignment and noise between the sketch and photo domains. (3) A new FG-SBIR dataset is introduced which has the biggest number of sketch-photo pairs for a single object category. Extensive experiments are carried out on three benchmarks. The results show that the proposed model significantly outperforms the state-of-the-art and both proposed novel components contribute to the superior performance.

3. Methodology

3.1. Overview

The architecture of the proposed model is illustrated in Fig. 2. It is a Siamese network with three CNN branches, corresponding to a query sketch, a positive photo and a negative photo respectively. The positive-negative relation can be defined by the matching relationship, e.g., if the true match photo is the positive, any false match can be used as the negative. Alternatively, if the sketches and photos are annotated explicitly by similarity, relative similarity ordering can be used as supervision information. The CNNs extract deep features from the three input images and feed them to a triplet ranking loss to enforce the ranking order (positive should be closer to the query than the negative using the extracted feature). With the learned model, for a given query sketch s and a set of M candidate photos $\{p_j\}_{j=1}^{M} \in \mathcal{P}$, we need to compute the similarity between s and p_j and use it to rank the set of gallery photos in the hope that the true match for the query sketch is ranked at the top. The similarity measure is computed by the high order distance function (detailed later), based on the domain invariant representations $F_\theta(\cdot)$ produced by the three Siamese CNN branches.

Similar to [46], the CNN base net is the Sketch-a-Net [47] which was originally designed for sketch recognition. We follow the same data preprocessing step to extract edge maps from each photo image to narrow the domain gap. The model is also pretrained on sketch recognition and category level SBIR data following exactly the same procedure as in [46]. The key differences are (1) an added attention module, (2) coarse-fine fusion, and (3) HOELF loss, which will described in details in the following sections.

3.2. Attention Modelling

A soft attention paradigm is adopted. Given a feature map computed at any convolutional layer of a CNN, a soft attention module will take it as input and generate an attention mask. This mask then used to re-weight the input feature map to get an attended feature map which is fed into the next layer of the network. In our model, the attention module is added to the output of the fifth convolutional+pooling

layer of the CNN in each branch (see Fig. 2).

We denote the input feature map as $\boldsymbol{f} \in \mathbb{R}^{H\times W\times C}$ where H and W are the filter map size and C is the number of feature channels. For the feature vector $f_{i,j} \in \mathbb{R}^C$ of the feature map at the spatial location (i, j), we can calculate its corresponding attention score $s_{i,j}$ by

$$\begin{aligned} s_{i,j} &= F_{att}(f_{i,j}; \boldsymbol{W}_a), \\ \alpha_{i,j} &= softmax(s_{i,j}), \end{aligned} \tag{1}$$

where $F_{att}(\cdot)$ is the mapping function learned by the attention module and $\boldsymbol{W_a}$ are the weights/parameters of the attention module. The final attention mask $\boldsymbol{\alpha} = [\alpha_{i,j}]$ is a probability map obtained by normalising the score matrix $\boldsymbol{s} = [s_{i,j}]$ using softmax. In our model, the attention module is a network consisting of two convolutional layers with kernel size 1. However, it can be replaced with any network. The attended feature map $\boldsymbol{f}^{att} = [f^{att}_{i,j}]$ is computed by element-wise product of the attention mask and the input feature map

$$f^{att}_{i,j} = \alpha_{i,j} \odot f_{i,j}. \tag{2}$$

Taking a conventional attention modelling approach, the attended feature map will be fed into the subsequent layer, which is FC6. However, due to the severe spatial misalignment of the query photo and either the positive or the negative photo, the attention mask could be somewhat imprecise and the resultant attended feature map $\boldsymbol{f}^{att}$ could be (a) corrupted by noise, and (b) loose any useful information in the original feature map $\boldsymbol{f}$. To overcome this problem, we introduce a shortcut connection architecture to link the input of the attention network directly to its output and combine them with an element-wise sum. The final attended feature map with shortcut connection is thus computed as

$$\boldsymbol{f}^{att}_s = \boldsymbol{f} + \boldsymbol{\alpha} \odot \boldsymbol{f}, \tag{3}$$

where '$+$' is element-wise sum and '$\odot$' is element-wise product. In this way, both the original feature map and the attended but imprecise feature map are combined and used as input to the next layer of the network.

3.3. Coarse-fine Fusion

Although the final attended feature map $\boldsymbol{f}^{att}_s$ is spatially aware and attentive to fine-grained details, these tend to be lost going through multiple subsequent fully connected layers, defeating the purpose of introducing attention modelling. To keep both the coarse and fine-grained information, a shortcut connection architecture is again employed here. Specifically, we fuse the attended feature map $\boldsymbol{f}^{att}_s$ with the output of the final FC layer (FC7) $\boldsymbol{f}^{FC7}$ to form the final feature representation $\boldsymbol{f}^{final}$ before it is fed into the loss layer (Fig. 2). A simple concatenation operation is used to fuse the two features. Before the fusion, we do global average pooling (GAP) on the attended feature map to reduce the dimension.

3.4. HOLEF Loss

Triplet Loss with a First-order Energy Function For a given triplet $t = (s, p^+, p^-)$ consisting of a query sketch s, a positive photo p^+ and a negative photo p^-, a conventional triplet ranking loss can be written as:

$$\begin{aligned} L_\theta\left(s, p^+, p^-\right) = \max(0, \Delta + D\left(F_\theta\left(s\right), F_\theta\left(p^+\right)\right) \\ - D\left(F_\theta\left(s\right), F_\theta\left(p^-\right)\right)), \end{aligned} \tag{4}$$

where θ are the parameters of the CNN with attention network, $F_\theta(\cdot)$ denotes the output of the corresponding network branch, *i.e.*, $\boldsymbol{f}^{final}$, Δ is the required margin of ranking for the hinge loss, and $D(\cdot, \cdot)$ denotes a distance between the two input representations, typically Euclidean distance. Considering $D(\cdot, \cdot)$ as a pairwise energy function, it is a first-order one due to the use of Euclidean distance which does element-wise subtraction of the feature. It does not consider the pairs of non-corresponding elements, thus implying alignment between the input feature representations, and not exploiting cross-channel correlation. It is thus particularly suboptimal, once we include the fine-grained attended feature map $\boldsymbol{f}^{att}_s$ in the feature representation.

Triplet Loss with a Higher-order Energy Function To compare two misaligned and noisy feature inputs, we can exploit higher order structural difference. To this end, we propose to compute a 2nd order feature difference using outer subtraction. Given two input feature vectors of k dimensions, the outer subtraction ($\ominus$) of the two is a $k \times k$ matrix. For example, when $k = 3$, we have:

$$\begin{aligned} F_\theta(s) \ominus F_\theta(p) &= \begin{bmatrix} {F_\theta}^1(s) \\ {F_\theta}^2(s) \\ {F_\theta}^3(s) \end{bmatrix} \ominus \begin{bmatrix} {F_\theta}^1(p) \\ {F_\theta}^2(p) \\ {F_\theta}^3(p) \end{bmatrix} \\ &= \begin{bmatrix} {F_\theta}^1(s) - {F_\theta}^1(p) & {F_\theta}^1(s) - {F_\theta}^2(p) & {F_\theta}^1(s) - {F_\theta}^3(p) \\ {F_\theta}^2(s) - {F_\theta}^1(p) & {F_\theta}^2(s) - {F_\theta}^2(p) & {F_\theta}^2(s) - {F_\theta}^3(p) \\ {F_\theta}^3(s) - {F_\theta}^1(p) & {F_\theta}^3(s) - {F_\theta}^2(p) & {F_\theta}^3(s) - {F_\theta}^3(p) \end{bmatrix} \end{aligned} \tag{5}$$

With outer subtraction, the difference between the elements at any position of the two input vectors are exhaustively computed, thus having the potential to deal with any form of feature misalignment.

With this outer-subtraction operator, we can design a 2nd order distance/energy-function based on the sum of the square of each element of the matrix. However, only a subset of these comparisons are expected to be useful, so we introduce a weighting factor to each element, resulting in the following energy function:

$$\mathcal{D}_H(F_\theta(s), F_\theta(p)) = \sum \left(F_\theta(s) \ominus F_\theta(p)\right)^{\circ 2} \odot \boldsymbol{W}, \tag{6}$$

where '$\circ 2$' is the element-wise square, and $\boldsymbol{W}$ is a $k \times k$ weight matrix. $\boldsymbol{W}$ is a learnable weighting layer matrix.

We can now replace the standard Euclidean loss in triplet ranking with our new energy function. Combined with appropriate regularisers, this leads to our high-order learnable energy function (HOELF) loss:

$$\begin{aligned} L_\theta\left(s, p^+, p^-\right) = \max(0, \Delta + \mathcal{D}_H\left(F_\theta\left(s\right), F_\theta\left(p^+\right)\right) \\ - \mathcal{D}_H\left(F_\theta\left(s\right), F_\theta\left(p^-\right)\right)) + \lambda\|\boldsymbol{W} - \boldsymbol{I}\|_1 \\ + \lambda\|\boldsymbol{W} - \boldsymbol{I}\|_F, \end{aligned} \tag{7}$$

where $\boldsymbol{I} \in \mathbb{R}^{k\times k}$ is the identity matrix and $||.||_F$ denotes the matrix Frobenious norm. Two regularisers are introduced in our loss. These elastic-net [50] style regularisers are designed to the keep $\boldsymbol{W}$ in the vicinity of $\boldsymbol{I}$, but prevent $\boldsymbol{W} - \boldsymbol{I}$ from being extremely sparse, i.e. $\boldsymbol{W}$ becoming diagonal and the HOLEF loss degenerating into a first-order loss. The weight λ for the two regularisers are set to 0.0005 in this work.

Ranking Score In the testing stage, given an query sketch s, the ranking score between the query sketch and each candidate photo p_i from a gallery set is computed as

$$\mathcal{R}_s(F_\theta(s), F_\theta(p_i)) = -\mathcal{D}_H(F_\theta(s), F_\theta(p_i)). \tag{8}$$

The rank scores are then used to rank the gallery set. The photo with the highest ranking score is the predicted match.

Alternative Higher-order Energy Function We are not aware any outer subtraction based higher-order energy function used as a loss for deep model training. However, outer product based ones are not uncommon. They have been used mainly for multi-view fusion, for example, fusing the text and image embeddings in visual question answering [21] and zero-shot recognition [4]. Outer product based distance is also used for formulating higher-order losses in Mahalanobis metric learning [41, 35]. Given two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, a Mahalanobis distance is defined as:

$$\begin{aligned} \mathcal{D}_M(\boldsymbol{x}, \boldsymbol{y}) &= (\boldsymbol{x} - \boldsymbol{y})^T \boldsymbol{M} (\boldsymbol{x} - \boldsymbol{y}) \\ &= \boldsymbol{x}^T \boldsymbol{M} \boldsymbol{x} + \boldsymbol{y}^T \boldsymbol{M} \boldsymbol{y} - 2\boldsymbol{x}^T \boldsymbol{M} \boldsymbol{y} \end{aligned} \tag{9}$$

where $\boldsymbol{M}$ is a learnable matrix. Compare Mahalanobis distance to the proposed distance in Eq. 6, it is clear that although both are 2nd order, there is a vital difference: In Mahalanobis distance, one first computes the element-wise subtraction $\boldsymbol{x} - \boldsymbol{y}$ and then the 2nd order bilinear product of the difference vectors. In other words, elements of different positions in the two vectors are not directly compared. It is thus not suitable for dealing with fine-grained feature misalignment and using correlation to compensate for noise in the sketch and photo feature vectors.

Figure 3. Examples of newly collected Handbag dataset.

4. Experiments

4.1. Datasets and Settings

Datasets We focus on the task of retrieving visually similar object instances from the same category – a setting resembling a real-world application where a customer searches for a specific product, e.g., shoe or handbag. Few FG-SBIR datasets are available publicly, and even fewer have more than 100 sketch-photo pairs from the same category to make the evaluation meaningful. We experiment on three datasets. **QMUL-Shoe** and **QMUL-Chair** from [46] contain 419 shoe and 297 chair sketch-photo pairs, respectively. The photos are real product photos collected from online shopping websites and the sketches are free-hand ones collected via crowdsourcing. We use 304 and 200 pairs for training and the rest for testing following the same splits as in [46]. There are 13,680 and 9,000 human triplet annotations which are used to train the triplet model. **Handbag** is a new dataset collected by us following similar protocol as the other two (photos from online catalogues and sketches crowd-sourced), resulting in 568 sketch-photo pairs. Handbags were specifically chosen to make the sketch-photo retrieval task more challenging, since handbags exhibit more complex visual patterns and have more deformable bodies than shoes and chairs. Among them, 400 are used for training and the rest for testing. The difference between this dataset and the other two is that we only have pairing information but not human triplet annotation. We thus generate the triplets using only true and false matches, rather than exhaustive similarity ranking. Following [46], we first extract edge maps from photos using the method of [49] and use them as input for the photo branch of our model. All images are resized to the same size of 256×256. Examples of the new Handbag dataset can be seen in Fig. 3.

Implementation Details Our model is implemented on TensorFlow. Each branch is pretrained in stages using a sketch recognition dataset and ImageNet photo-edgemap pairs, similarly to the procedure described in [46], before fine-tuning on each FG-SBIR dataset. The initial learning rate is 0.001 and the mini-batch size is 128. During training, we randomly crop a 225×225 sub-image as input and we do flipping with 0.5 probability. The atten-

QMUL-Shoe	acc.@1	acc.@10
HOG-BoW + rankSVM	17.39%	67.83%
Dense-HOG + rankSVM	24.35%	65.22%
ISN Deep + rankSVM	20.00%	62.61%
Triplet SN [46]*	52.17 %	92.17 %
Our model	**61.74%**	**94.78%**
QMUL-Chair	acc.@1	acc.@10
HOG-BoW + rankSVM	28.87%	67.01%
Dense-HOG + rankSVM	52.57%	93.81%
ISN Deep + rankSVM	47.42%	82.47%
Triplet SN [46]*	72.16 %	**98.96** %
Our model	**81.44%**	95.88%
Our Handbag	acc.@1	acc.@10
HOG-BoW + rankSVM	2.38%	10.71%
Dense-HOG + rankSVM	15.47%	40.48%
ISN Deep + rankSVM	9.52%	44.05%
Triplet SN [46]*	39.88%	82.14%
Our model	**49.40%**	**82.74%**

Table 1. Comparative results against baselines. '*' The results of Triplet SN [46] are the updated ones from their project webpage which are higher than the published results due to parameter retuning. The other baseline results are copied from [46] except those on Handbag, which are based on our own implementation.

tion module consists of 2 convolutional layers, both with kernel size 1×1. $\boldsymbol{W}$ in the HOLEF loss is learned as a trainable layer. A detailed description of the network architecture can be found in the Supplementary Material. Both our dataset and the trained model can be found at: http://sketchx.eecs.qmul.ac.uk/downloads/.

4.2. Comparative Results

Baselines Four baseline models are chosen for comparison. Two are hand-crafted feature based models, namely **HOG-BoW+RankSVM** and **Dense-HOG+RankSVM**. HOG features are classic for sketch-recognition [16] and SBIR [10] problem and it is the most commonly used hand-crafted feature before the popularity of deep features. Dense HOG is obtained by concatenating HOG features over a dense grid. A RankSVM model is used with the features to compute the final ranking score. Among the other two baseline models, **ISN Deep+RankSVM** uses the deep features extracted from Sketch-a-Net [47], which was trained for sketch recognition. The prior state of the art model **Triplet SN** was the first end-to-end deep model for SBIR [46]. It has an identical base network architecture as ours and differs in the lack of attention model and the use of conventional first-order Euclidean triplet loss.[1]

Results We use the ratio of correctly predicting the true match at top-1 and at top-10 (acc.@1 and acc.@10) as the evaluation metrics. The performance of all compared models are reported in Table 1. The results suggest that (1) The two end-to-end learned deep models are clearly superior to the other baselines. (2) The proposed model significantly outperforms all baseline models on all three datasets. The improvement is particularly clear at top-1 – around 9% increase in top-1 accuracy is obtained on all three datasets against the second best model. For each query sketch, there are typically a handful of visually very similar photos; the lower-rank accuracy, especially at top-1, thus is a better indication on how well the model is capable of distinguishing fine-grained subtle differences between candidate photos. Note that the drop of acc.@10 on Chair dataset can be explained by the introduction of the attention module. With attention, our model is able to focus on discriminative local parts. Yet, very occasionally the attention module locates the wrong parts which happen to be shared by other objects with globally very different appearance. This problem is more acute for chair than shoe and handbag because part sharing across different sub-categories is more common.

[1] Further experimental results on the recently released Sketchy database [31] can be found in Supplementary Materials.

QMUL-Shoe	acc.@1	acc.@10
Base	52.17%	92.17%
Base + CFF	58.26%	93.04%
Base + HOLEF	56.52%	88.70%
Full: Base + CFF + HOLEF	**61.74%**	**94.78%**
QMUL-Chair	acc.@1	acc.@10
Base	72.16%	**98.96%**
Base + CFF	79.38%	95.88%
Base + HOLEF	74.23%	97.94%
Full: Base + CFF + HOLEF	**81.44%**	95.88%
Our Handbag	acc.@1	acc.@10
Base	39.88%	82.14%
Base + CFF	48.21%	83.33%
Base + HOLEF	40.48%	**83.93%**
Full: Base + CFF + HOLEF	**49.40%**	82.74%

Table 2. Contributions of the different components.

4.3. Ablation Study

Contributions of each Component We have introduced two novel components in our model: the coarse-fine fusion (CFF) to combine the attended convolutional feature map with the final FC layer output and the HOLEF loss. In order to evaluate the contributions of each component, we compare our full model (**Full: Base+CFF+HOLEF**) with three stripped-down versions: baseline model with coarse-fine fusion (**Base+CFF**), baseline model with HOLEF loss (**Base+HOLEF**) and baseline without either (**Base**) which becomes the Triplet SN model [46]. Table 2 shows clearly that each novel component improves the base model and

QMUL-Shoe	with attention	without attention
Base	54.78%	52.17%
Base + CFF	58.26%	57.39%
Base + HOLEF	57.39%	56.52%
Our model	61.74%	58.26%

QMUL-Chair	with attention	without attention
Base	74.23%	72.16%
Base + CFF	79.38%	75.25%
Base + HOLEF	75.26%	74.23%
Our model	81.44%	77.32%

Our Handbag	with attention	without attention
Base	41.07%	39.88%
Base + CFF	48.21%	47.02%
Base + HOLEF	40.48%	40.48%
Our model	49.40%	48.21%

Table 3. Effectiveness of the attention module (acc.@1).

QMUL-Shoe	with shortcut	without shortcut
Base + attention	54.78%	15.65%
Base + CFF	58.26%	26.96%
Our model	61.74%	27.83%

QMUL-Chair	with shortcut	without shortcut
Base + attention	74.23%	39.18%
Base + CFF	79.38%	48.45%
Our model	81.44%	49.48%

Our Handbag	with shortcut	without shortcut
Base + attention	41.07%	17.26%
Base + CFF	48.21%	24.40%
Our model	49.40%	23.81%

Table 4. Effect of shortcut connection in attention module (acc.@1).

when both are combined we achieved the best performance indicating that they are complementary to each other.

Contributions of the Attention Module Two experiments are carried out. First, we evaluate how effective our attention module is, not only to the final full model, but also to the various stripped-down versions. Table 3 show that almost invariantly each model variant benefits from having an attention module to locate the most discriminative part of the object to compare across the two domains. Second, we evaluate the usefulness of the proposed shortcut connection architecture in the attention module which is designed to deal with the potentially imprecise attention mask caused by spatial misalignment between the compared sketch and photo pair. Table 4 shows that having this shortcut connection architecture is vital: without the shortcut, i.e., having a conventional soft attention module, the attended feature map on its own is too noisy to be useful.

HOELF vs. other Alternative Triplet Losses To further validate the effectiveness of our HOLEF loss, we compare with: (i) conventional triplet loss with Euclidean distance, (ii) triplet loss with weighted Euclidean distance, and (iii) triplet loss with Mahalanobis distance. The first two are first order whilst the third is second order. The last two have learnable weights while the first does not. All models have the same base network and attention model as well as CFF. They thus differ only in the loss used. The results are

QMUL-Shoe	acc.@1	acc.@10
Triplet loss with Euclidean	58.26%	93.04%
Triplet loss with Weighted Euclidean	58.26%	93.04%
Triplet loss with Mahalanobis	52.17%	89.57%
Our HOLEF	**61.74**%	**94.78**%

QMUL-Chair	acc.@1	acc.@10
Triplet loss with Euclidean	79.38%	95.88%
Triplet loss with Weighted Euclidean	79.38%	95.88%
Triplet loss with Mahalanobis	78.35%	95.88%
Our HOLEF	**81.44**%	**95.88**%

Our Handbag	acc.@1	acc.@10
Triplet loss with Euclidean	48.21%	**83.33**%
Triplet loss with Weighted Euclidean	48.81%	82.14%
Triplet loss with Mahalanobis	44.64%	79.76%
Our HOLEF	**49.40**%	82.74%

Table 5. Comparison on different losses.

Figure 4. Visualisation of attention masks of sample photo-sketch pairs in all three categories.

Figure 5. Comparison of the retrieval results of our model and Triplet SN [46]. For each example, the top row is our retrieval result with attention mask superimposed on the query sketch, and the bottom row is retrieval result of the same sketch using Triplet SN.

shown in Table 5. It can be seen that: (1) The proposed 2nd order outer subtraction based HOLEF loss is the best. (2) Even with learnable weights, both weighted Euclidean and Mahalanobis distance in most cases cannot beat the conventional triplet loss with Euclidean distance. (3) Even with a 2nd order energy function, the bilinear product of element-wise subtraction used in Mahalonobis distance is ineffective at dealing with the noise and feature misalignment of the two domains.

4.4. Visualisation and Qualitative Results

Attention Processing In Fig. 4 we offer visualisations of the attention maps learned using our model. It can be seen that: (i) Across all three datasets, attention tends to be associated with salient parts of the object having complicated and distinct visual pattern, e.g., shoelaces, wheels on chairs, and bag buckles. (ii) Attention masks seem to align well across sketch and photo domains, e.g., the cross pattern on the back of the chair.

Qualitative Retrieval Results We further provide qualitative examples of our retrieval results in Fig. 5, compared with those obtained using Triplet SN [46]. We observe that our spatial-semantic attention model is better at disambiguating subtle visual details. For example, on the first shoe example (left), attending to the shoelace region resulting in the correct shoe being retrieved as Rank 1. Similarly on bags, attending to the stripe pattern resulted in the correct bag being returned amongst bags whose overall shapes are almost identical. For the sofa on the right, despite both models returning the correct top-1 match, our attended model was able to filter out the sofa bed which was ranked 2nd by Triplet SN.

5. Conclusion

We have proposed a novel deep spatial-semantic attention model for FG-SBIR. By introducing attention modelling and shortcut connections, it is able to concentrate on the subtle differences between local regions of a sketch and photo images and compute deep features containing both fine-grained and high-level semantics. However, fine-grained noise and cross-domain feature channel misalignment challenge energy functions for cross-domain matching. We therefore introduced a novel HOLEF loss to make the model robust against this. The effectiveness of the proposed model has been validated by extensive experiments.

References

[1] Y. Cao, C. Wang, L. Zhang, and L. Zhang. Edgel index for large-scale sketch-based image search. In *CVPR*, 2011. 2

[2] Y. Cao, H. Wang, C. Wang, Z. Li, L. Zhang, and L. Zhang. Mindfinder: interactive sketch-based image search on millions of images. In *ACM-MM*, 2010. 2

[3] M. Eitz, J. Hays, and M. Alexa. How do humans sketch objects? In *TOG*, 2012. 1

[4] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013. 5

[5] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, 2016. 2

[6] L. A. Gatys, A. S. Ecker, and M. Bethge. Texture synthesis and the controlled generation of natural stimuli using convolutional neural networks. *CoRR*, 2015. 1

[7] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *ECCV*, 2014. 3

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 3

[9] R. Hu, M. Barnard, and J. Collomosse. Gradient field descriptor for sketch based retrieval and localization. In *ICIP*, 2010. 2

[10] R. Hu and J. Collomosse. A performance evaluation of gradient field HOG descriptor for sketch based image retrieval. *CVIU*, 2013. 1, 2, 6

[11] R. Hu, T. Wang, and J. Collomosse. A bag-of-regions approach to sketch based image retrieval. In *ICIP*, 2011. 2

[12] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015. 2

[13] S. James, M. Fonseca, and J. Collomosse. Reenact: Sketch based choreographic design from archival dance footage. In *ICMR*, 2014. 2

[14] B. Klare, Z. Li, and A. K. Jain. Matching forensic sketches to mug shot photos. *PAMI*, 2011. 1

[15] Y. Li, T. M. Hospedales, Y.-Z. Song, and S. Gong. Fine-grained sketch-based image retrieval by matching deformable part models. In *BMVC*, 2014. 1, 2

[16] Y. Li, T. M. Hospedales, Y.-Z. Song, and S. Gong. Free-hand sketch recognition by multi-kernel feature learning. *CVIU*, 2015. 6

[17] T. Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *ICCV*, 2015. 2

[18] Y. Lin, C. Huang, C. Wan, and W. Hsu. 3D sub-query expansion for improving sketch-based multi-view image retrieval. In *ICCV*, 2013. 2

[19] Y. Liu, Y. Guo, and M. S. Lew. On the exploration of convolutional fusion networks for visual recognition. In *MMM*, 2017. 3

[20] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *arXiv preprint arXiv:1612.01887*, 2016. 2

[21] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 2016. 5

[22] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *CVPR*, 2015. 1

[23] E. Mathias, H. Kristian, B. Tamy, and A. Marc. An evaluation of descriptors for large-scale image retrieval from sketched feature lines. *Computers & Graphics*, 2010. 2

[24] E. Mathias, H. Kristian, B. Tamy, and A. Marc. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *TVCG*, 2011. 1, 2

[25] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *NIPS*, 2014. 2

[26] H. Nam, J.-W. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. *arXiv preprint arXiv:1611.00471*, 2016. 2

[27] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 3

[28] S. Ouyang, T. M. Hospedales, Y.-Z. Song, and X. Li. Forgetmenot: Memory-aware forensic facial sketch matching. In *CVPR*, 2016. 1

[29] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015. 3

[30] X. Ren. Multi-scale improves boundary detection in natural images. In *ECCV*, 2008. 3

[31] P. Sangkloy, N. Burnell, C. Ham, and J. Hays. The sketchy database: Learning to retrieve badly drawn bunnies. In *SIGGRAPH*, 2016. 1, 2, 3, 6

[32] R. G. Schneider and T. Tuytelaars. Sketch classification and classification-driven analysis using fisher vectors. *TOG*, 2014. 1

[33] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 3

[34] P. Sermanet, A. Frome, and E. Real. Attention for fine-grained categorization. *arXiv preprint arXiv:1412.7054*, 2014. 2

[35] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li. Embedding deep metric for person re-identification: A study against large variations. In *ECCV*, 2016. 3, 5

[36] J. Song, Y.-Z. Song, T. Xiang, T. Hospedales, and X. Ruan. Deep multi-task attribute-driven ranking for fine-grained sketch-based image retrieval. In *BMVC*, 2016. 1, 3

[37] 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. 2, 3

[38] C. Wang, Z. Li, and L. Zhang. Mindfinder: image search by interactive sketching and tagging. In *WWW*, 2010. 2

[39] F. Wang, L. Kang, and Y. Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *CVPR*, 2015. 1, 2

[40] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *CVPR*, 2014. 3

[41] K. Weinberger and L. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 2009. 3, 5

[42] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *CVPR*, 2015. 2

[43] S. Xie and Z. Tu. Holistically-nested edge detection. In *ICCV*, 2015. 3

[44] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 2

[45] S. Yang and D. Ramanan. Multi-scale recognition with dag-cnns. In *ICCV*, 2015. 3

[46] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, and C.-C. Loy. Sketch me that shoe. In *CVPR*, 2016. 1, 2, 3, 5, 6, 8

[47] Q. Yu, Y. Yang, Y.-Z. Song, T. Xiang, and T. Hospedales. Sketch-a-net that beats humans. *BMVC*, 2015. 1, 3, 6

[48] H. Zhang, S. Liu, C. Zhang, W. Ren, R. Wang, and X. Cao. Sketchnet: Sketch classification with web images. In *CVPR*, 2016. 2

[49] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 5

[50] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. 5