DLA-Net for FG-SBIR: Dynamic Local Aligned Network for Fine-Grained Sketch-Based Image Retrieval

Jiaqing Xu
jqxu@bupt.edu.cn
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications

Haifeng Sun*
hfsun@bupt.edu.cn
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications

Qi Qi
qiqi8266@bupt.edu.cn
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications

Jingyu Wang*
wangjingyu@bupt.edu.cn
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications

Ce Ge
cege.research@gmail.com
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications

Lejian Zhang
zhanglejian@ebupt.com
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications

Jianxin Liao
liaojx@bupt.edu.cn
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications

ABSTRACT

Fine-grained sketch-based image retrieval is considered as an ideal alternative to keyword-based image retrieval and image search by image due to the rich and easily accessible characteristics of sketches. Previous works always follow a paradigm that first extracting image global feature with convolution neural network and then optimizing the model with triplet loss. Many efforts on narrowing the domain gap and extracting discriminating features are made by these works. However, they ignored that the global feature is not good at capturing fine-grained details. In this paper, we emphasize the local features are more discriminating than global feature in FG-SBIR and explore an effective way to utilize local features. Specifically, Local Aligned Network (LA-Net) is proposed first, which solves FG-SBIR by directly aligning the mid-level local features. Experiment manifests it can beat all previous baselines and is easy to implement. LA-Net is hoped to be a new strong baseline for FG-SBIR. Next, Dynamic Local Aligned Network (DLA-Net) is proposed to enhance LA-Net. The question of spatial misalignment caused by the abstraction of the sketch is not considered by LA-Net. To solve this question, a dynamic alignment mechanism is introduced into LA-Net. This new mechanism makes the sketch interact with the photo and dynamically decide where to align according to the different photos. The Experiment indicates DLA-Net successfully addresses the question of spatial misalignment. It gains a significant performance boost over LA-Net and outperforms the state-of-the-art in FG-SBIR. To the best of our knowledge, DLA-Net is the first model that beats humans on all datasets—QMUL FG-SBIR, QMUL Handbag, and Sketchy.

CCS CONCEPTS

• **Computing methodologies → Visual content-based indexing and retrieval**.

KEYWORDS

FG-SBIR, LA-Net, DLA-Net, sketch, local features, dynamic alignment mechanism

ACM Reference Format:
Jiaqing Xu, Haifeng Sun, Qi Qi, Jingyu Wang, Ce Ge, Lejian Zhang, and Jianxin Liao. 2021. DLA-Net for FG-SBIR: Dynamic Local Aligned Network for Fine-Grained Sketch-Based Image Retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475705

1 INTRODUCTION

The free-hand sketches, as their rich and easily accessible characteristics, have been a common form of expression since ancient

[*] Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8651-7/21/10...$15.00
https://doi.org/10.1145/3474085.3475705

times. With the popularity of touch-screen devices, there is an increasing attention on sketch-related issues, including sketch recognition [6] [41] [13], sketch to image synthesis [2] [22] [17], sketch segmentation [38] [42], and sketch based image retrieval [3] [40] [31] [32]. As a sub-issue of sketch-based image retrieval, fine-grained sketch-based image retrieval (FG-SBIR) [42] [40] [31] [19] is gaining more and more attention due to its commercial value in the field of image retrieval.

The problem FG-SBIR tries to address is finding the most similar photo based on the input sketch. It faces three challenges. First, there is a large domain gap between sketches and photos as sketches are made up of sparse black lines, while photos are made up of dense color pixels. Second, sketches are highly abstract. As shown in Figure 1a, different drawing levels and different comprehension of drawers lead to a large intra-class difference. Last but not least, fine-grained details are crucial for FG-SBIR. There may be only a very slight difference between the query sketch and unmatched photos as shown in Figure 1b. It is challenging to obtain the representation of fine-grained details.

Existing FG-SBIR models have proposed various methods to solve these three challenges. Early works [31] [39] mainly focused on the first and second challenges. Triplet network with triplet loss was used in these works to narrow the domain gap. Meanwhile, the powerful representational capability of CNNs made these models robust to the abstraction of the sketch. Recent works [40] [27] [26] paid more attention to the third challenge. Attention mechanism, generative task, and jigsaw puzzle self-supervision task are introduced by these methods to better encode fine-grained information. However, all of these methods utilized the global feature extracted by global average pooling (GAP) [20] or fully connected layer. The global feature is not discriminating enough as it only focuses on notable parts while ignores the subtle fine-grained details. The notable parts are crucial to solve the coarse-grained task like image classification. But for a fine-grained task like FG-SBIR, subtle fine-grained details are more important. This means the local features may be more suitable for FG-SBIR.

To this end, we lay emphasis on local features and explore an effective way to solve FG-SBIR by utilizing local features. We first propose a simple but strong baseline—Local Aligned Network (LA-Net). LA-Net takes images as an aggregate of fine-grained details and considers FG-SBIR as the process of finding sketch and photo matching pairs that share most fine-grained details. It extracts local features from the mid-level layer of the backbone network first and then directly aligns the local features by narrowing the distance of features at the same spatial location.We next propose Dynamic Local Aligned Network (DLA-Net) to better utilize local features. LA-Net assumes that the paired sketch and photo are strictly spatial aligned. However, due to the abstraction of sketches, it is common to find spatial misalignment between the paired sketch and photo. To solve this question, a dynamic alignment mechanism is introduced. This mechanism makes the query sketch interact with gallery photos and dynamically decide where to align.

Our contributions are as follows:

- We emphasize the importance of local features and propose a new baseline named Local Aligned Network (LA-Net). LA-Net solves FG-SBIR by aligning local features.
- We propose Dynamic Local Aligned Network (DLA-Net) to better utilize local features. DLA-Net introduces a dynamic alignment mechanism to solve the spatial misalignment that LA-Net can not handle.
- Extensive experiments show LA-Net surpasses previous baselines on three FG-SBIR datasets, which demonstrates local features are more discriminating than global feature in FG-SBIR.
- Extensive experiments demonstrate dynamic alignment mechanism is beneficial for the usage of local features. DLA-Net outperforms existing approaches by a significant margin. Furthermore, to the best of our knowledge, DLA-Net first beats humans on all three datasets.

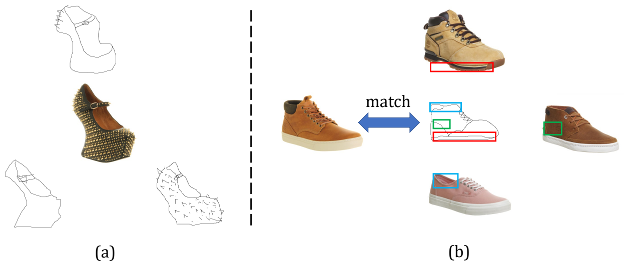

Figure 1: (a) An example of abstraction of sketch. (b) An example of fine-grained characteristic of FG-SBIR, different colored boxes represent the dissimilar regions.

2 RELATED WORK

2.1 Category-level SBIR

The goal of category-level SBIR is to find a photo, which has the same category as the query sketch. This is one of the earliest studied sketch-related problems. Early approaches are mainly based on handcrafted descriptors, including HOG[10] [9], color histogram[14], BOW[7], etc. Later works focus on deep learning methods[1] [37] [36]. These methods use CNNs with triplet loss or contrastive loss to narrow the domain gap. Recently, a more difficult problem—zero-shot SBIR has become the mainstream issue in category-level SBIR.

2.2 Fine-grained SBIR

FG-SBIR further requires the photo and sketch match at the instance level. Compared with category-level SBIR, this is a relatively new problem. FG-SBIR was first proposed in [18], which employed a deformable-part model to tackle this problem. Subsequent FG-SBIR models mainly used deep learning methods. The two earlier models [31] [39] were similar. They all employed triplet networks with triplet loss to learn a unified embedding feature space. A classification loss is additionally added in [31] to fit the multi-category datasets. These two works did not pay attention to fine-grained details. To solve this problem, recent works proposed various methods. Pang *et al.* [26] introduced a generative task to make the embedding feature space more sensitive towards fine-grained details. In [40], an attention mechanism was introduced to make the model pay more attention to fine-grained details. A mixed modal jigsaw puzzle pre-training task is proposed in [27] to extract more effective features for FG-SBIR. Sain *et al.* [30] studied the hierarchical trait of sketches and learned a more discriminating feature by propagating across different detail levels. Although these methods made progress in FG-SBIR, they all had one question that all of them used global feature, which can not capture enough fine-grained details. In this paper, our DLA-Net learns to encode fine-grained details by dynamically aligning local features.

2.3 Global and local features in CNNs

Since the introduction of AlexNet [16] in 2012, CNNs have become the main research method in the field of computer vision due to their powerful representational capability. CNNs were first used in image classification tasks, in which the abstract global feature is more effective. Besides, many commonly used pre-trained models are also pre-trained for classification tasks. So, it is common to utilize global feature when using CNNs in other tasks. However, the global feature extracted by fully connected layer or global average pooling layer [20] loses the fine-grained information. This makes the local features are more effective than global feature in fine-grained related tasks, such as object detection [29], fine-grained image classification [12] [21], and image retrieval [24] [5]. For FG-SBIR, fine-grained details are also very important, but there is no work that has attempted to utilize local features. Our DLA-Net makes use of the local features and the outstanding performance proves that the local features are more suitable for FG-SBIR.

3 LA-NET

Given the query sketches and gallery photos, most previous methods take the input image as a whole and used the Euclidean distance between global features as a similarity. These methods may ignore many fine-grained details due to the use of the global feature. Different from them, we take images as an aggregate of fine-grained details and considers FG-SBIR as the process of finding sketch and photo matching pairs that share most fine-grained details. Following this assume, LA-Net is proposed. As shown in Figure 2, LA-Net first uses Local Feature Extractor to obtain the mid-level local feature map and then directly align the local features in Local Aligned Module.

3.1 Local Feature Extractor

Similar to previous works, LA-Net also uses triplet networks. Meanwhile, considering the inter-domain differences, the sketch branch and photo branch of LA-Net do not share parameters. Local Feature Extractor can take any CNNs as the backbone. In this paper, we employ ResNet50 [8] as the backbone due to its competitive performance and concise architecture. In order to obtain the local features, GAP layer and the following layers are removed. It is worth noting that we also remove the final convolution layer and use the output feature map of conv3 layer. We do this to better encode fine-grained details. It is widely studied in [20] [23] that the deeper layer leads higher semantic level and a bigger receptive field. To encode fine-grained details which usually exist in a relatively small area, the mid-level local features with lower semantic level and smaller receptive field are better. Formally, given a triplet (S, P^+, P^-) as input, in which S represents query sketch, P^+ represents positive photo, P^- represents negative photo. Local Feature Extractor will output the mid-level feature map $\mathbf{f} \in \mathbb{R}^{C\times H\times W}$, where $[H, W]$ is the spatial size and C is the dimension of the feature map.

3.2 Local Aligned Module

Given the output of Local Feature Extractor, Local Aligned Module first uses local L2 normalization to normalize the input mid-level feature map. Normalizing image feature with L2 normalization is a shared method in FG-SBIR because it can stabilize the training process by restricting the embedding feature space to hypersphere. However, simply normalizing the feature map will make the local feature at every spatial location be polluted by others. Thus, a local L2 normalization is proposed in LA-Net to normalize every local feature independently. Specifically, given feature map $\mathbf{f} \in \mathbb{R}^{C\times H\times W}$, the local L2 normalization is defined as:

$$\mathbf{g}(x,y) = \frac{\mathbf{f}(x,y)}{max(\|\mathbf{f}(x,y)\|_2, \epsilon)} \tag{1}$$

$$\|\mathbf{f}(x,y)\|_2 = \sqrt{\sum_{i=1}^{C} f_i^2(x,y)} \tag{2}$$

where ϵ is a small value to avoid division by zero.

Then, Local Aligned Module calculates the distance between features of sketch and photo. It first calculates the local distance, which is the euclidean distance between a pair of local features at the same location on the feature map of sketch and photo. These local distances are viewed as the similarity between the localized areas of sketch and photo. For paired sketches and photos, there should be more similar localized areas. It means the value of most local distances should be small. We use 2-norm to calculate the total distance because 2-norm tends to make every input element small. Formally, the total distance between a sketch and a photo is formulated as:

$$d_i(S,P) = \sqrt{\sum_{j=1}^{C}[g_j^s(x,y) - g_j^p(x,y)]^2} \tag{3}$$

$$D(S,P) = \sqrt{\sum_{i=1}^{H\times W} d_i^2(S,P)} \tag{4}$$

Finally, LA-Net is optimized using triplet loss, given a triplet (S, P^+, P^-) as input, triplet loss is defined as:

$$L_{tri} = max(0, \Delta + D(S,P^+) - D(S,P^-)) \tag{5}$$

where Δ is the margin between paired features and unpaired features.

4 DLA-NET

4.1 Limitation of LA-Net

Local distance in LA-Net is calculated on the same location of sketch and photo, which means LA-Net assumes that there is a strict spatial alignment between sketches and photos. If the sketch is well-drawn, this assumption is reasonable. However, due to the abstraction of sketches, it is a common phenomenon to see spatial misalignment between paired sketch and photo. As shown in Figure 3a, the buckle part of the sketch and photo are spatially misaligned. This misalignment will puzzle the model and weaken the ability of

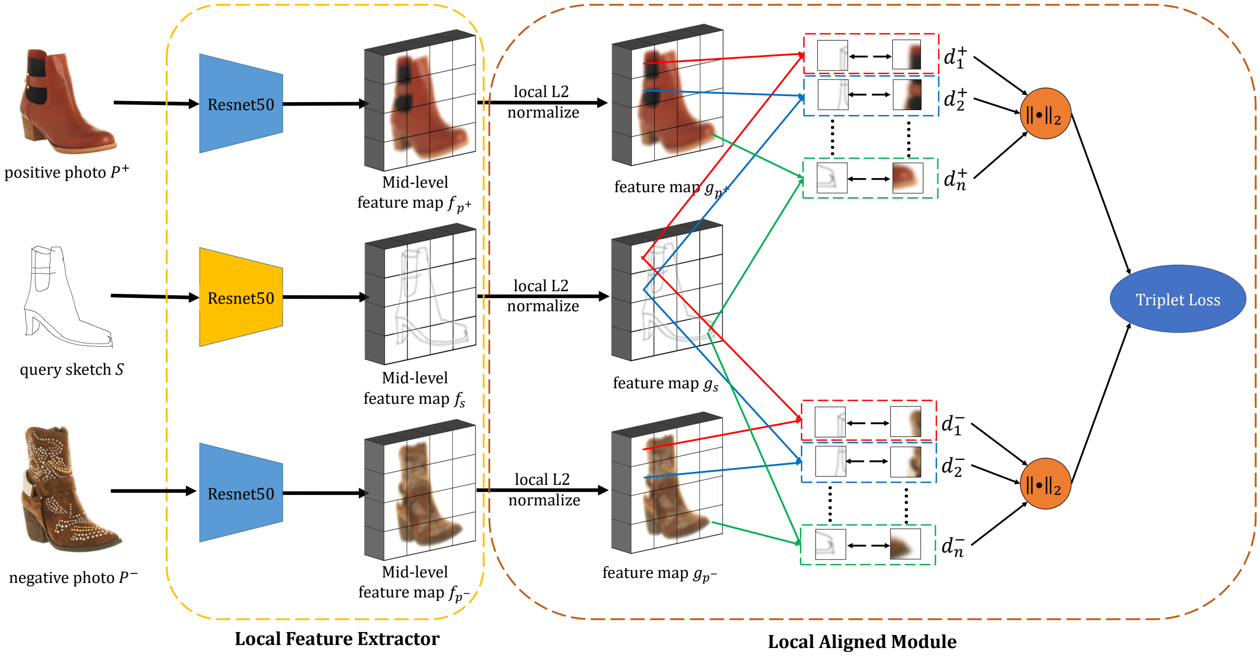

Figure 2: Architecture of LA-Net.

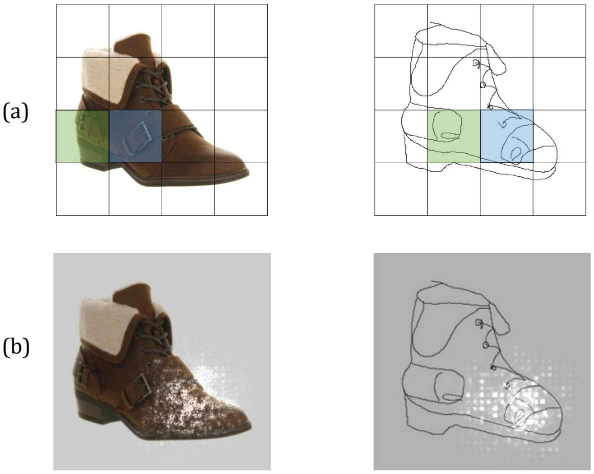

Figure 3: (a) Visualization of the spatial misalignment between paired sketch and photo. The same color block represents the area that should be aligned. (b) Visualization of the area local distance focus using guided backpropagation.

the model to learn discriminating local features. To further demonstrate the effect of spatial misalignment, we also make a visualization to visualize the area that local distance focuses on. Specifically, we utilize guided backpropagation [34] on one of the local distances of the paired sketch and photo. The visualization result is shown in Figure 3b. It can be observed that the area LA-Net focuses on is semantically mismatch when there is a spatial misalignment.

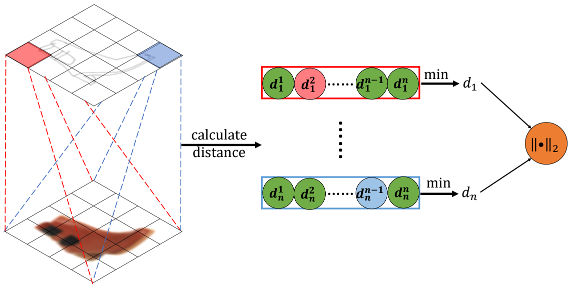
calculate distance
d_1^1 d_1^2 ...... d_1^{n-1} d_1^n
min
d_1
d_n^1 d_n^2 ...... d_n^{n-1} d_n^n
min
d_n
$\|\bullet\|_2$

Figure 4: Illustration of dynamic alignment mechanism. Local features of sketch slide over the whole feature map of photo and calculate local distances. The minimum local distances of every local feature of sketch are summed up using 2-norm to obtain the total distance.

4.2 Dynamic Alignment Mechanism

To solve the spatial misalignment, we improve Local Aligned Module and get a new model named Dynamic Local Aligned Network (DLA-Network). A dynamic alignment mechanism that is similar to convolution operation is introduced. As shown in Figure 4, every local feature of sketch slides over the whole feature map of photo and dynamically chooses the aligned local feature depending on different photos. This mechanism makes the sketch interact with the photo, which is consistent with human behavior. When performing FG-SBIR task, people always switch between sketches and photos to find the matching and unmatching areas. Sain *et al.* [30] first introduced cross-modal interaction in FG-SBIR. However, the sketch and photo features are merged in their cross-modal interaction. We think it is unreasonable to merge the two features that need to be compared later. So our DLA-Net only finds the aligned local feature instead of merging sketch and photo features. Specifically, the difference between LA-Net and DLA-Net is the calculation method of local distance. New local distance is changed as below:

$$d_i^k(S,P) = \sqrt{\sum_{j=1}^{C}[g_j^s(x,y) - g_j^p(x^k,y^k)]^2} \tag{6}$$

$$d_i(S,P) = min(d_i^1(S,P), d_i^2(S,P), \cdots, d_i^{W\times H}(S,P)) \tag{7}$$

5 EXPERIMENTS

5.1 Datasets and Settings

Three datasets are used in our experiments:

- **QMUL FG-SBIR** dataset consists of three subdataset, namely **QMUL-Shoe-v1**, **QMUL-Chair-v1**,**QMUL-Shoe-v2**. There are 419 and 297 sketch-photo pairs in QMUL-Shoe-v1 and QMUL-Chair-v1, respectively. 304 and 200 pairs are used to train and rest to test as mentioned in [39]. QMUL-Shoe-v2 is an extended version of QMUL-Shoe-v1. The number of photos is extended to 2000. Each photo has three or more hand-drawn sketches lead to 6730 sketches in total. We follow the split in [40], which uses 1800 pairs to train and rest to test.
- **QMUL Handbag** dataset was introduced in [33]. This dataset is similar to QMUL-Shoe-v1 and QMUL-Chair-v1. It contains 568 sketch-photo pairs. The split is consistent with [33] which uses 400 pairs to train and rest to test. We also do not use the human triplet annotations in this dataset.
- **Sketchy** dataset is the largest SBIR dataset. It is a multi-category dataset, which contains 74,425 sketches and 12,500 gallery photos spanning 125 categories. Each category has 100 photos and 5 or more corresponding sketches for each photo. There are 10 pairs in each category to test and the rest to train. This split is as same as [31].

Evaluation Metric. The goal of FG-SBIR is to find the matching photo. To evaluate the retrieval accuracy, acc.@1 is commonly used in previous methods. This metric is the percentage of sketches whose top-1 retrieval result is the true-match photo. We also use this metric. Multi-view testing is a commonly used testing strategy in previous methods [39] [33] [40]. TC-Net [19] has proven this strategy can boost some methods with the computational burden increasing in the testing stage. Considering the large computational burden, we do not use the multi-view testing strategy.

5.2 Implementation Details

Training. Our method is implemented on the Pytorch platform. As mentioned in Section 3.1, we employ ImageNet pre-trained ResNet50 model as the backbone feature extractor and remove the layers after conv3 layer (For Sketchy dataset, these layers are reserved to complete classification task as [31]). The sketch branch and photo branch do not share parameters and the shape of output feature map is 1024 × 16 × 16. The margin Δ in Eq. 5 is set to 0.1. Vanilla triplet loss is hard to optimize, we use the batch all triplet loss [4] when implementing the model. We set the batch size to 32 and train the model using Adam [15] optimiser. The model is trained for 100 epochs with learning rate are set as 0.0001 and 0.00001 for QMUL datasets and Sketchy dataset, respectively. Besides, for Sketchy dataset, the weight of triplet loss and classification loss is set to 48 and 1, respectively.

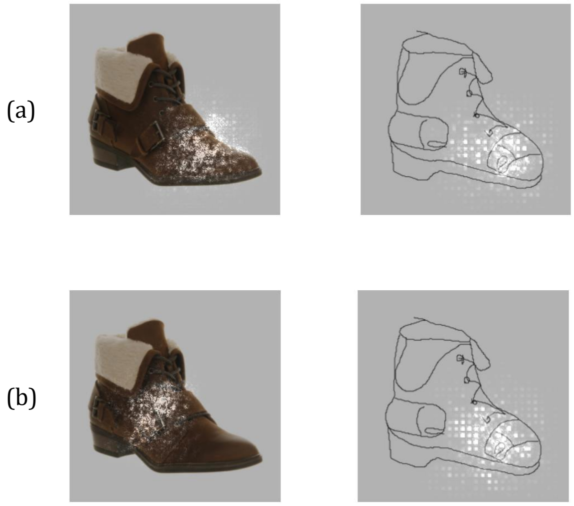

Figure 5: Visualization of the area that local distance focus. (a) Visualization of LA-Net. (b) Visualization of DLA-Net.

Data Processing. There are two views on whether to use edge map. TripletSN [39] and DSSA [33] convert photo to edge map to narrow the domain gap, while Sketchy [31] and TC-Net [19] do not use edge map, due to the fact it is expensive and unstable to train the network by edge maps. We follow the latter to use the original photos. The input sketches and photos are first resized to 288 × 288 and then are randomly cropped to 256 × 256. It is worth noting that we do not use random horizontal flip to augment datasets because we hold the view that FG-SBIR should have pose sensitivity.

5.3 Competitors

Deep learning methods are proven to be better than hand-crafted methods in [19] [40]. Thus, we only compare our model with deep learning methods. The competitors are split into two groups: (1)Baselines: **TripletSN** [39] uses Sketch-a-Net [41] as feature extractor and optimized the model using triplet loss. **Sketchy** [31] is similar to TripletSN except that it uses GoogLeNet [35] as backbone and adds a classification loss for Sketchy dataset. **Triplet-ResNet** replaces the backbone of TripletSN with ResNet50. We compare with this baseline to eliminate the effect of different backbones. (2) Benchmarks: **DSSA** [33] improves TripletSN using attention mechanism and higher-order energy function. **CGL** [26] adds a generative task to preserve all the domain invariant information that

Table 1: Comparative results on FG-SBIR specific datasets.

Method		QMUL-Shoe-v1	QMUL-Chair-v1	QMUL-Handbag	QMUL-Shoe-v2	Sketchy
Baselines	TripletSN [39]	52.17%	72.16%	39.90%	30.93%	36.72%
	Sketchy [31]	-	-	-	-	37.10%
	Triplet-ResNet	28.70%	62.89%	37.39%	35.89%	40.91%
Benchmarks	DSSA [33]	61.74%	81.44%	49.40%	-	-
	CGL [26]	-	-	-	-	50.14%
	DSM [28]	54.80%	85.60%	51.20%	-	-
	CMHM [30]	-	-	-	36.27%	-
	SMJP [27]	56.52%	85.98%	62.97%	36.52%	53.45%
	TC-Net [19]	63.48%	95.88%	-	40.02%	40.81%
Ours	LA-Net	57.39%	93.81%	56.52%	42.34%	43.13%
	DLA-Net	**79.13%**	**98.97%**	**69.57%**	**50.15%**	**54.92%**
	Human	76.52%	94.85%	50%	49.50%	54.27%

Table 2: Comparison of using different level local features.

Method	QMUL-Shoe-v1	QMUL-Chair-v1	QMUL-Handbag	QMUL-Shoe-v2	Sketchy
Triplet-ResNet	28.70%	62.89%	37.39%	35.89%	40.91%
LA-Net	**57.39%**	**93.81%**	**56.52%**	**42.34%**	43.13%
LA-Net-high	41.74%	86.60%	48.70%	41.29%	**50.41%**
DLA-Net	**79.13%**	**98.97%**	**69.57%**	**50.15%**	**54.92%**
DLA-Net-high	41.74%	78.35%	46.09%	42.94%	49.54%

Table 3: Comparison of different normalization strategy.

Method	QMUL-Shoe-v1	QMUL-Chair-v1	QMUL-Handbag	QMUL-Shoe-v2	Sketchy
LA-Net (global L2)	55.65%	**93.81%**	51.30%	37.54%	42.64%
LA-Net (w/o L2)	56.52%	92.78%	52.17%	42.19%	38.87%
LA-Net	**57.39%**	**93.81%**	**56.52%**	**42.34%**	**43.13%**
DLA-Net (global L2)	54.78%	**98.97%**	62.61%	43.39%	53.70%
DLA-Net (w/o L2)	66.09%	97.94%	66.09%	48.50%	53.39%
DLA-Net	**79.13%**	**98.97%**	**69.57%**	**50.15%**	**54.92%**

is useful for cross-domain reconstruction. **DSM** [28] cast shape matching as metric learning with CNNs. It first turned images into edge maps and then utilized CNNs to extract image descriptors. **CMHM** [30] extracts features by parsing the hierarchy of sketches and reinforces features via cross-modal interaction. **TC-Net** [19] uses DenseNet-169 [11] as backbone and introduces auxiliary classification loss to facilitate the network. **SMJP** [27] uses mixed-modal jigsaw puzzle to replace the ImageNet pre-train procedure, which can produce more suitable feature for FG-SBIR. **Human baselines** are also very important baselines. We report human baselines to prove the effectiveness of our DLA-Net. The human baselines of QMUL FG-SBIR dataset and QMUL Handbag dataset are reported in [40], while the human baseline of Sketchy is reported in [31].

5.4 Performance Analysis

The comparative results are shown in Table 1. We make the following observations:

LA-Net is a new strong baseline. Comparing with the previous baselines—TripletSN and Sketchy, LA-Net obtains the best retrieval accuracy on all datasets. It is worth noting that even with the use of multi-view testing and complex pre-training, TripletSN is also surpassed by LA-Net. LA-Net only makes a slight modification on TripletSN and is trained without bells and whistles. We hope it will serve as a strong baseline for FG-SBIR.

Local features are more suitable for FG-SBIR. We compare LA-Net with Triplet-ResNet, which is similar to LA-Net except utilizing the global feature. LA-Net outperforms Triplet-ResNet by a remarkable margin on all datasets. This demonstrates that the discriminating ability of local features is stronger than the global feature. Thereby, local features are more suitable for FG-SBIR.

Dynamic alignment mechanism improves LA-Net significantly. Although LA-Net already gains a relatively high retrieval accuracy, DLA-Net also gains a significant improvement on LA-Net. As mentioned in Section 4, dynamic alignment mechanism is proposed to solve the problem of spatial misalignment. To demonstrate this, we visualize where the local distance focus using guided backpropagation [34]. The visualization results are shown in Figure 5. It can be observed that LA-Net focuses on the areas which

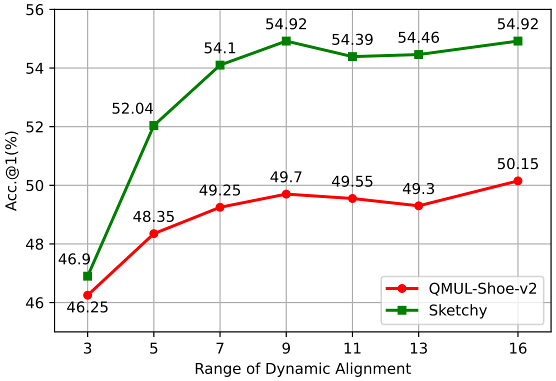

Figure 6: Impact of different range of dynamic alignment. These variants are tested on QMUL-Shoe-v2 dataset and Sketchy dataset.The range of 16 means sliding over the whole feature map.

are located at the same location of paired sketch and photo, while DLA-Net focuses on the areas that have the same semantics. This demonstrates dynamic alignment mechanism can effectively solve the problem of spatial misalignment, which is the key to the improvement of DLA-Net.

Comparison with benchmarks. Comparing with the previous benchmarks, DLA-Net surpasses all previous models. More importantly, DLA-Net beats humans on all datasets for the first time. This further demonstrates the effectiveness of DLA-Net.

5.5 Ablation Study

In this section, we compare LA-Net and DLA-Net with several variants to validate some key design choices.

The benefit of mid-level local features. As mentioned in Section 3.1, LA-Net uses the mid-level local features instead of the high-level local features, because we take the attitude that mid-level local features are more suitable to encode fine-grained details. To verify the validity of mid-level local features, we compare Triplet-ResNet, LA-Net, and DLA-Net with LA-Net-high and DLA-Net-high. These two variants use the high-level local features which are the output of the conv4 layer of ResNet50. The result is shown in Table 2. We make the following observations: (1) Both LA-Net and LA-Net-high perform better than Triplet-ResNet, which indicates local features are more discriminating in FG-SBIR. (2) Comparing with mid-level local features, high-level local features have a bad effect on both models except LA-Net on Sketchy dataset. This demonstrates that the mid-level local features are more beneficial to FG-SBIR than high-level local features. For the exceptions on Sketchy dataset, the more severe spatial misalignment of Sketchy dataset may be the cause. As the CNNs going deeper, receptive field will become bigger, which is beneficial for spatial misalignment. Besides, comparing with QMUL FG-SBIR and QMUL Handbag in which photos only have white background, Sketchy photos have complicated background. This leads to considerable noise. High-level feature can effectively eliminate this noise, which also makes LA-Net-high performs better than LA-Net. (3) Comparing the results of LA-Net-high and DLA-Net-high, it is clear to see that dynamic alignment mechanism have a tiny or even bad effect when using high-level local features. This means the mid-level local features are necessary for dynamic alignment mechanism.

Table 4: Comparative results on CC-FG-SBIR

Method	acc.@1
Triplet-ResNet50	12.68%
[25]	22.60%
LA-Net	20.00%
DLA-Net	27.41%

The benefit of local L2 normalization. We compare local L2 normalization with two different normalization strategies—global L2 normalization and without L2 normalization. Global L2 normalization normalizes all local features and without L2 normalization simply abandons normalization. These two different normalization strategies are tested on LA-Net and DLA-Net. Table 3 suggests global L2 normalization and without L2 normalization performs differently on different datasets, but they all worse than local L2 normalization. This indicates local L2 normalization is vital for FG-SBIR.

The area size of dynamic alignment. Dynamic alignment mechanism solve the question of spatial misalignment by dynamically choosing aligned local feature. The area size of dynamic alignment determines the ability of DLA-Net to solve this question. We test different area sizes of dynamic alignment on QMUL-Shoe-v2 dataset and Sketchy dataset. Figure 6 shows the retrieval accuracy improves as the area size increases until reaching the size of 9. After that, DLA-Net reaches a stable performance. This result is reasonable because the spatial misalignment only exists in a relatively small area in FG-SBIR. It means we do not need to slide over the whole photo feature map. However, when implementing dynamic alignment mechanism using a small area size, a selective mask is needed which will reduce computing efficiency. Thus, we set the area size of dynamic alignment to 16 in DLA-Net.

5.6 Qualitative Visualization

Figure 7 shows some retrieved results on QMUL-Shoe-v2 dataset using Triplet-ResNet, LA-Net, and DLA-Net, respectively. From the first row, we can observe that LA-Net, and DLA-Net focus more on fine-grained details and are sensitive to the location of details. This further demonstrates local features are more suitable for FG-SBIR. The second row shows the case that query sketches are more abstract. These results show LA-Net is unable to cope with the serious abstract sketches, while DLA-Net can retrieval the correct results. This indicates the dynamic alignment mechanism is good at tackling the abstraction of sketch.

To further demonstrate the ability of DLA-Net to solve the abstraction of sketch, we also make a visualization to find the aligned area of paired sketch and photo. Note that we only visualize the discriminative areas as shown in Figure 8. These results show DLA-Net is able to find the correct local area by using dynamic alignment mechanism.

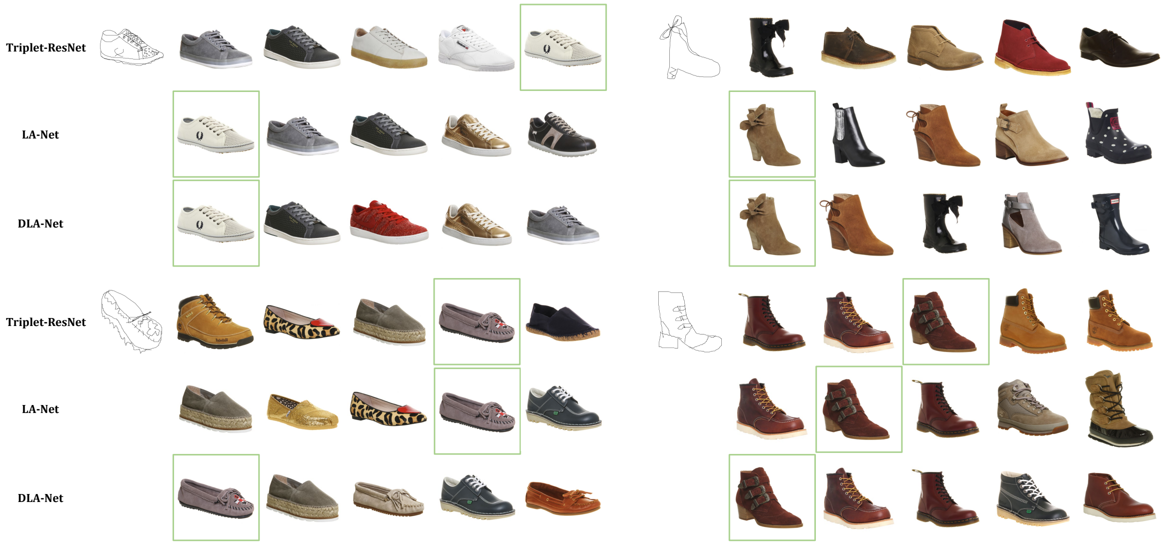

Figure 7: Visualization of the retrieved results on QMUL-Shoe-v2 dataset for different models, the correct results are marked with green boxes.

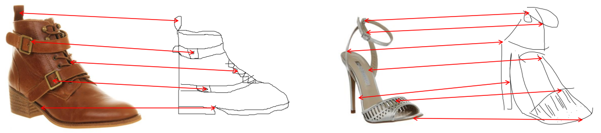

Figure 8: Visualization of the aligned area learned by DLA-Net.

6 FURTHER ANALYSIS

Due to the fact that sketches are difficult to obtain, cross-category fine-grained sketch-based image retrieval (CC-FG-SBIR) is proposed in recent years [27] [25]. We note that these methods also used the global feature. The global feature is of high semantic level which leads it is unfavorable for CC-FG-SBIR. While, mid-level local features focus more on fine-grained details, which are likely to be shared by other categories. This makes it may be more suitable for CC-FG-SBIR. Thus, we validate the generalization capability of our LA-Net and DLA-Net. We perform experiment on Sketchy dataset. Following the same data split as [25], we split Sketchy dataset into 104 train and 21 test categories, in which the test categories are not present in ImageNet-1K. The training epoch is also set to 20 as same as [25]. Comparative results are shown in Table 4. It is clear to see LA-Net and DLA-Net perform better than Triplet-ResNet50. Furthermore, our DLA-Net surpasses the model specifically designed for CC-FG-SBIR. This indicates DLA-Net has powerful generalizing capability.

To evaluate the computational complexity of DLA-Net, we conduct an additional experiment. On the QMUL-Shoe-v2 dataset, Triplet-ResNet50 needs 32s and 3s per epoch for train and test, respectively. LA-Net needs 28s and 5s, while DLA-Net needs 33s and 7s. It can be seen that the computational complexity of DLA-Net is higher than others because the sketch interacts with photos. However, comparing with the significant improvement in retrieval accuracy, this cost is relatively small and acceptable. We will focus on the question of computational complexity in future work.

7 CONCLUSION

This paper demonstrates the local features are more discriminating in FG-SBIR and explores an effective way to utilize local features. Firstly, we propose a new strong baseline for FG-SBIR named LA-Net. LA-Net takes the image as an aggregate of fine-grained details and directly aligns the mid-level local features. With a simple architecture, LA-Net can surpass all previous baselines. This proves local features, especially the mid-level local features are more suitable for FG-SBIR. Then, we note that directly aligning local features can not solve the question of spatial misalignment. Thus, we propose DLA-Net which introduces a dynamic alignment mechanism into LA-Net. DLA-Net achieves the best performance and even beats humans on three FG-SBIR specific datasets, which demonstrates dynamic alignment mechanism is an effective way to utilize local features. Besides, we also find DLA-Net performs best on CC-FG-SBIR, which further proves the effectiveness of DLA-Net.

ACKNOWLEDGMENTS

This work was supported in part by the National Key R&D Program of China 2020YFB1807805, National Natural Science Foundation of China under Grants (62071067, 62001054 and 61771068), in part by the National Postdoctoral Program for Innovative Talents under Grant BX20200067, Ministry of Education and China Mobile Joint Fund MCM20200202, in part by China Postdoctoral Science Foundation under Grant 2021M690469, in part by BUPT innovation and entrepreneurship support program 2021-YC-T021.

REFERENCES

[1] Tu Bui, Leonardo Sampaio Ferraz Ribeiro, Moacir Ponti, and John P. Collomosse. 2018. Sketching out the details: Sketch-based image retrieval using convolutional neural networks with multi-stage regression. *Comput. Graph.* 71 (2018), 77–87. https://doi.org/10.1016/j.cag.2017.12.006

[2] Wengling Chen and James Hays. 2018. SketchyGAN: Towards Diverse and Realistic Sketch to Image Synthesis. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 9416–9425. https://doi.org/10.1109/CVPR.2018.00981

[3] Sounak Dey, Pau Riba, Anjan Dutta, Josep Lladós, and Yi-Zhe Song. 2019. Doodle to Search: Practical Zero-Shot Sketch-Based Image Retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2179–2188. https://doi.org/10.1109/CVPR.2019.00228

[4] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. 2015. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognit.* 48, 10 (2015), 2993–3003. https://doi.org/10.1016/j.patcog.2015.04.005

[5] Mihai Dusmanu, Ignacio Rocco, Tomás Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. 2019. D2-Net: A Trainable CNN for Joint Description and Detection of Local Features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 8092–8101. https://doi.org/10.1109/CVPR.2019.00828

[6] Mathias Eitz, James Hays, and Marc Alexa. 2012. How do humans sketch objects? *ACM Trans. Graph.* 31, 4 (2012), 44:1–44:10. https://doi.org/10.1145/2185520.2185540

[7] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. 2011. Sketch-Based Image Retrieval: Benchmark and Bag-of-Features Descriptors. *IEEE Trans. Vis. Comput. Graph.* 17, 11 (2011), 1624–1636. https://doi.org/10.1109/TVCG.2010.266

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778. https://doi.org/10.1109/CVPR.2016.90

[9] Rui Hu, Mark Barnard, and John P. Collomosse. 2010. Gradient field descriptor for sketch based retrieval and localization. In *Proceedings of the International Conference on Image Processing, ICIP 2010, September 26-29, Hong Kong, China*. IEEE, 1025–1028. https://doi.org/10.1109/ICIP.2010.5649331

[10] Rui Hu and John P. Collomosse. 2013. A performance evaluation of gradient field HOG descriptor for sketch based image retrieval. *Comput. Vis. Image Underst.* 117, 7 (2013), 790–806. https://doi.org/10.1016/j.cviu.2013.02.005

[11] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2261–2269. https://doi.org/10.1109/CVPR.2017.243

[12] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. 2016. Part-Stacked CNN for Fine-Grained Visual Categorization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 1173–1182. https://doi.org/10.1109/CVPR.2016.132

[13] Qi Jia, Xin Fan, Meiyu Yu, Yuqing Liu, Dingrong Wang, and Longin Jan Latecki. 2020. Coupling Deep Textural and Shape Features for Sketch Recognition. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (Eds.). ACM, 421–429. https://doi.org/10.1145/3394171.3413810

[14] Toshikazu Kato, Takio Kurita, Nobuyuki Otsu, and Kyoji Hirata. 1992. A sketch retrieval method for full color image database-query by visual example. In *11th IAPR International Conference on Pattern Recognition, ICPR 1992. Conference A: Computer Vision and Applications, The Hague, Netherlands, August 30-September 3, 1992*. IEEE, 530–533. https://doi.org/10.1109/ICPR.1992.201616

[15] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger (Eds.). 1106–1114. https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[17] Yuhang Li, Xuejin Chen, Binxin Yang, Zihan Chen, Zhihua Cheng, and Zheng-Jun Zha. 2020. DeepFacePencil: Creating Face Images from Freehand Sketches. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (Eds.). ACM, 991–999. https://doi.org/10.1145/3394171.3413684

[18] Yi Li, Timothy M. Hospedales, Yi-Zhe Song, and Shaogang Gong. 2014. Intra-category sketch-based image retrieval by matching deformable part models. In *British Machine Vision Conference, BMVC 2014, Nottingham, UK, September 1-5, 2014*, Michel François Valstar, Andrew P. French, and Tony P. Pridmore (Eds.). BMVA Press. http://www.bmva.org/bmvc/2014/papers/paper112/index.html

[19] Hangyu Lin, Yanwei Fu, Peng Lu, Shaogang Gong, Xiangyang Xue, and Yu-Gang Jiang. 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, MM 2019, Nice, France, October 21-25, 2019*, Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi (Eds.). ACM, 1676–1684. https://doi.org/10.1145/3343031.3350900

[20] Min Lin, Qiang Chen, and Shuicheng Yan. 2014. Network In Network. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1312.4400

[21] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. 2015. Bilinear CNN Models for Fine-Grained Visual Recognition. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 1449–1457. https://doi.org/10.1109/ICCV.2015.170

[22] Runtao Liu, Qian Yu, and Stella X. Yu. 2020. Unsupervised Sketch to Photo Synthesis. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 12348)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 36–52. https://doi.org/10.1007/978-3-030-58580-8_3

[23] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard S. Zemel. 2016. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). 4898–4906. https://proceedings.neurips.cc/paper/2016/hash/c8067ad1937f728f51288b3eb986afaa-Abstract.html

[24] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. 2018. LF-Net: Learning Local Features from Images. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). 6237–6247. https://proceedings.neurips.cc/paper/2018/hash/f5496252609c43eb8a3d147ab9b9c006-Abstract.html

[25] Kaiyue Pang, Ke Li, Yongxin Yang, Honggang Zhang, Timothy M. Hospedales, Tao Xiang, and Yi-Zhe Song. 2019. Generalising Fine-Grained Sketch-Based Image Retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 677–686. https://doi.org/10.1109/CVPR.2019.00077

[26] Kaiyue Pang, Yi-Zhe Song, Tony Xiang, and Timothy M. Hospedales. 2017. Cross-domain Generative Learning for Fine-Grained Sketch-Based Image Retrieval. In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*. BMVA Press. https://www.dropbox.com/s/5lrdcpy1i7va2ce/0122.pdf?dl=1

[27] Kaiyue Pang, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, and Yi-Zhe Song. 2020. Solving Mixed-Modal Jigsaw Puzzle for Fine-Grained Sketch-Based Image Retrieval. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 10344–10352. https://doi.org/10.1109/CVPR42600.2020.01036

[28] Filip Radenovic, Giorgos Tolias, and Ondrej Chum. 2018. Deep Shape Matching. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V (Lecture Notes in Computer Science, Vol. 11209)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer, 774–791. https://doi.org/10.1007/978-3-030-01228-1_46

[29] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 6 (2017), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

[30] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. 2020. Cross-Modal Hierarchical Modelling for Fine-Grained Sketch Based Image Retrieval. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press. https://www.bmvc2020-conference.com/assets/papers/0102.pdf

[31] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. 2016. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Trans. Graph.* 35, 4 (2016), 119:1–119:12. https://doi.org/10.1145/2897824.2925954

[32] Yuming Shen, Li Liu, Fumin Shen, and Ling Shao. 2018. Zero-Shot Sketch-Image Hashing. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 3598–3607. https://doi.org/10.1109/CVPR.2018.00379

[33] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. 2017. Deep Spatial-Semantic Attention for Fine-Grained Sketch-Based Image Retrieval. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 5552–5561. https://doi.org/10.1109/ICCV.2017.592

[34] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. 2015. Striving for Simplicity: The All Convolutional Net. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6806

[35] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 1–9. https://doi.org/10.1109/CVPR.2015.7298594

[36] Xinggang Wang, Xiong Duan, and Xiang Bai. 2016. Deep sketch feature for cross-domain image retrieval. *Neurocomputing* 207 (2016), 387–397. https://doi.org/10.1016/j.neucom.2016.04.046

[37] Peng Xu, Yongye Huang, Tongtong Yuan, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, Zhanyu Ma, and Jun Guo. 2018. SketchMate: Deep Hashing for Million-Scale Human Sketch Retrieval. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 8090–8098. https://doi.org/10.1109/CVPR.2018.00844

[38] Lumin Yang, Jiajie Zhuang, Hongbo Fu, Kun Zhou, and Youyi Zheng. 2020. SketchGCN: Semantic Sketch Segmentation with Graph Convolutional Networks. *CoRR* abs/2003.00678 (2020). arXiv:2003.00678 https://arxiv.org/abs/2003.00678

[39] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen Change Loy. 2016. Sketch Me That Shoe. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 799–807. https://doi.org/10.1109/CVPR.2016.93

[40] Qian Yu, Jifei Song, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. 2021. Fine-Grained Instance-Level Sketch-Based Image Retrieval. *Int. J. Comput. Vis.* 129, 2 (2021), 484–500. https://doi.org/10.1007/s11263-020-01382-3

[41] Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. 2017. Sketch-a-Net: A Deep Neural Network that Beats Humans. *Int. J. Comput. Vis.* 122, 3 (2017), 411–425. https://doi.org/10.1007/s11263-016-0932-3

[42] Changqing Zou, Qian Yu, Ruofei Du, Haoran Mo, Yi-Zhe Song, Tao Xiang, Chengying Gao, Baoquan Chen, and Hao Zhang. 2018. SketchyScene: Richly-Annotated Scene Sketches. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV (Lecture Notes in Computer Science, Vol. 11219)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer, 438–454. https://doi.org/10.1007/978-3-030-01267-0_26