Fine-Grained Image Retrieval: the Text/Sketch Input Dilemma

Fine-Grained Image Retrieval: the Text/Sketch Input Dilemma

Jifei Song[1]
j.song@qmul.ac.uk
Yi-Zhe Song[1]
yizhe.song@qmul.ac.uk
Tao Xiang[1]
t.xiang@qmul.ac.uk
Timothy Hospedales[1, 2]
t.hospedales@ed.ac.uk

[1] SketchX Research Lab
Queen Mary University of London
London, UK

[2] School of Informatics
University of Edinburgh
Edinburgh, UK

Abstract

Fine-grained image retrieval (FGIR) enables a user to search for a photo of an object instance based on a mental picture. Depending on how the object is described by the user, two general approaches exist: sketch-based FGIR or text-based FGIR, each of which has its own pros and cons. However, no attempt has been made to systematically investigate how informative each of these two input modalities is, and more importantly whether they are complementary to each thus should be modelled jointly. In this work, for the first time we introduce a multi-modal FGIR dataset with both sketches and sentences description provided as query modalities. A multi-modal quadruplet deep network is formulated to jointly model the sketch and text input modalities as well as the photo output modality. We show that on its own the sketch modality is much more informative than text and each modality can benefit the other when they are modelled jointly.

1 Introduction

Fine-grained image retrieval (FGIR) [19, 29, 33] aims to search for photos containing specific object instances. It presents a paradigm shift to conventional image retrieval tasks, by offering instance-level retrieval that underpins the need for many commercial applications such as searching an online shopping website product catalogue. Specifically, different to traditional image retrieval paradigms where input queries and results are often coarse (*e.g.*, keywords and general object categories), FGIR aims to retrieve specific object instances based on a user's precise description. Such a description can be provided in two very different forms: text and sketch.

Text being a conventional input modality is arguably the most intuitive – people have got used to typing in keywords in search engines to retrieve text documents. Keyword-based text query can also do a decent job for category-level image retrieval. For example, using the keyword 'shoe' in a Google/Bing image search engine generates very satisfactory results - the first few return pages all contain shoe images. However, when it comes to instance-level

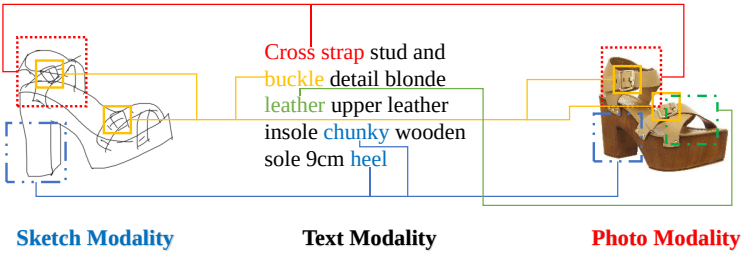

Figure 1: Relationship between sketch, photo, and text modalities.

or FGIR, using text as an input modality is problematic: it is good at describing semantic concepts or attributes of the objects but weak in detailing spatial layout and complex shape related characteristics. After all, one picture is worth a thousand words. A user can write a sentence in a pinch but will not be bothered with writing an essay for retrieving a photo.

This limitation of text as an input modality for FGIR has inspired a recent surge of interest in sketch-based FGIR approaches [19, 33]. Human sketches have been advocated by many as a natural input modality since it implicitly captures both fine-grained appearance and holistic structure information [14, 19, 32, 33]. A sketch is perhaps worth one hundred words but takes much less efforts to produce. With the popularity of touchscreens, drawing sketch has never been easier. However, sketch-based retrieval paradigms still suffer the major drawback of varying drawing skills amongst users, which ultimately render it unintuitively for many – the '*I can't sketch*' response is common. On top of that, certain visual characteristics can be cumbersome to sketch, yet straightforward to describe in text (*e.g.*, material and fine texture). It is thus natural to hypothesise that these two input modalities are complementary to each other (see Figure 1 for an example) and thus should be modelled jointly. Nevertheless, as far as we know, there is no systematic study on how these two modalities fare in FGIR and importantly, how their complementarity can be exploited so that even when a single modality is used during testing, it can still benefit from a joint modelling process during training.

In this paper, we set out to answer the question whether text or sketch as an input modality is a clear favourite when it comes to fine-grained retrieval of photos, or if there is complementary information to be explored for them to benefit from each other – and if there is, how it can be exploited in a joint model? The first contribution of this work is to provide the first dataset for FGIR with both sketch and text as query modalities. Specifically, each object instance has three modalities: photo, sketch and sentence description enabling research into not only sketch-text based FGIR in this work, but also fine-grained retrieval tasks between any of the three modalities.

As the second contribution, we propose a multi-modal quadruplet deep network to align sketch, text and photo embeddings. The main novelty is a quadruplet loss after the final FC layers of the network, which not only aligns the three modalities, but also provides fine-grained ranking similar to triplet losses previously used in two-modality fine-grained retrieval [19, 33]. As the final contribution, we carry out extensive experiments to investigate the usefulness of each modality as an input query on its own and when combined with other modalities. We demonstrate that on its own the sketch modality is far more informative than text even when multiple sentences are used, but both sketch and text benefit from being modelled jointly during training, even when used as the sole query modality during testing.

2 Related work

Fine-grained Sketch-based Image Retrieval The traditional category-level SBIR problem has been well studied [2, 9, 15, 24, 27, 35]. It was not until very recently that unique characteristics of sketches had been advocated for fine-grained retrieval. Deep neural network (DNN) provides an end-to-end solution for fine-grained SBIR [17, 19, 21, 22, 33], and are proven to be superior to shallow methods based on hand-crafted features [12, 13]. [33] formulates a triplet-ranking network to align sketch and photo modalities. It adopts a Siamese architecture, and utilise a triplet loss to learn a joint embedding space. The Sketchy network [19] uses a Heterogeneous architecture instead and employ GoogleNet on each branch to learn modality dependent feature representations. It follows that a triplet loss is used on the final FC layers to align the two modalities. A classification loss is also used after both sketch and photo branches to ensure that the retrieval result belongs to the correct category. In this paper, we adopt the former Siamese architecture, where weights between sketch branch and photo branch are shared, and extend the triplet ranking loss to a quadruplet one in order to embed three modalities (sketch, text and photo).

Towards Fine-grained Text-based Image Retrieval There exists plenty of work on learning a text-photo embedding space for image search [7, 23, 30], captioning [4, 11, 23] and visual question answering [5, 31]. However, to the best of our knowledge, no prior work had specifically addressed the fine-grained retrieval problem using text as input. This is not surprising since as previously discussed text can become tedious and imprecise when descriptions become more detailed (i.e., from keywords to full textual descriptions), rendering them not the most intuitive for fine-grained retrieval. One line of work investigates using image captions to retrieve images. For example, a convolution neural network (CNN) learns a visual embedding which is then fed into a recurrent neural network (RNN) caption generator, to form a CNN-RNN architecture [23]. However, this line of work typically studies captions to describe the layout of objects in a scene image, rather than text to describe the fine-grained characteristics of one object. We directly compare with [23] in our experiments, and show that our three-way embedding can significantly boost text-photo retrieval performance.

Multi-modal Learning [25] offers an excellent survey of the field, including shallow methods and deep learning methods. For the shallow methods, an effective and representative approach is to learn a multi-view embedding with canonical correlation analysis (CCA) to align the images, tags and keywords, and then conduct the multi-view retrieval [7]. For the deep learning methods, a typical algorithm will at first learn the latent representation for each modality by a certain deep network, like auto encoder. After that, ranking loss after feature layer or correlation score from CCA layer is applied to supervise the deep multi-modal framework [26, 33]. In the inference stage, cross-similarity metric is built on the two or more cross-modality latent representations, and guide the cross-modal retrieval [1, 16]. However, the multi-modal learning on fine-grained instance-level under the same category is still not well investigated. In our work, we use the proposed quadruplet loss to align the learned representation for instance-level sketch, photo and text modalities, and more importantly, to exploit the fine-grained discriminative power across the embeddings from different modalities.

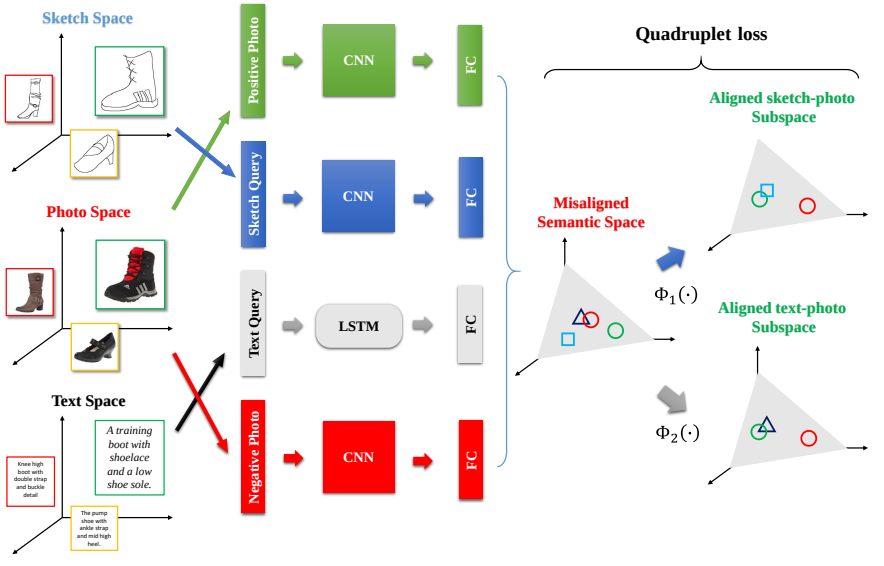

Figure 2: Architecture of the proposed multi-modal learning framework. For the Quadruplet loss visualisation: the light blue square indicates the anchor sketch, dark blue triangle the anchor text, and the green and red circles are for the positive and negative photos, respectively.

3 Methodology

3.1 Network Architecture

The architecture of our model is shown in Figure 2. It consists of four branches and extends the common architecture of a triplet ranking network: the middle two branches encodes sketch (S) and text (T) respectively, whereas the top and bottom branches are standard positive (PP) and negative (NP) photos branches as per the triplet fine-grained network of [33]. Each of the three branch configurations of our multi-modal framework, S-PN-NN, and T-PN-NN, relates to the task of fine-grained sketch-based image retrieval and fine-grained text-based image retrieval respectively. A novel quadruplet loss unifies these two related tasks and aligns these cross modality embeddings.

3.2 Fine-grained Sketch and Photo Feature Embedding

In the fine-grained sketch-based image retrieval task, it is important to learn a deep representation which encodes the fine-grained visual features shared between sketch and photo modalities. To achieve this, branches similar to the state-of-the-art sketch-photo ranking model in [33] are constructed, where we use Siamese convolution and pooling layers with weights tied among different domains. Then fully-connected (FC) layers are applied to reduce the high dimensional convolution layer feature to a lower dimensional feature space, while the following FC layer project the embedding from each modality to shared latent space.

We also apply the pairwise dropout strategy [6] for the activation in Siamese branches.

This is to depress the negative influence of standard dropout strategy on learning the ranking on pairwise/triplet feature map, as different masks will introduce mask difference error when we compare features in the ranking loss. Our experiments show that this pairwise dropout strategy is helpful on multi-view matching/ranking tasks.

3.3 Fine-grained Text Description Embedding

It is natural to utilize RNN based methods as our language model, to exploit the high-level information embedded in text descriptions. We use bidirectional long short-term memory (LSTM) network to capture the fine-grained text features since it gives the best performance amongst alternatives. In detail, sentences are chunked to tokenized word lists, then words are fed into into a word embedding (learned based on the gensim model [18]). The encoded vectors with timestamps are then sent to the bidirectional LSTM network to train the cells. The LSTM cells are then updated following [8]. Finally, different to sentence generation or image caption models, where they use the output from all word units, under a per-word softmax loss to predict each word, we only take the last hidden activation of the bidirectional LSTM [28] as the overall representation of the input sentence, *i.e.*, the text embedding.

3.4 Multi-modal Alignment

Given the learned sketch-photo and text-photo embeddings, the following task is to align these cross-modal embeddings. A cross-modal quadruplet loss is proposed to align the different embeddings. Given an instance quadruplet sample $\{s,t,p^+,p^-\}$, where the s, t, p^+, p^-, representing the anchor sketch, anchor text, positive photo and negative photo, respectively, the multi-modal model is supervised by our multi-modal quadruplet loss as below,

$$\begin{aligned} L\left(s,t,p^+,p^-\right) = &\max\left(0,\Delta+D\left(\Phi_1(f(s)),\Phi_1(f(p^+))\right)-D\left(\Phi_1(f(s)),\Phi_1(f(p^-))\right)\right) \\ &+\max\left(0,\Delta+D\left(g\left(t\right),\Phi_2(f(p^+))\right)-D\left(g(t),\Phi_2(f(p^-))\right)\right) \end{aligned} \tag{1}$$

, where $g(t)$, $f(s)$, $f(p^+)$ and $f(p^-)$ denotes the learned anchor text, anchor sketch, the positive and negative photo embedding, respectively. $D(\cdot,\cdot)$ is the distance metric, here we take the squared euclidean distance of l_2 normalised features to measure the cross-domain similarity. The margin in the quadruplet loss is Δ. Two linear transform layers are embedded in our quadruplet loss as $\Phi_1(\cdot)$ and $\Phi_2(\cdot)$ to further adapt domain gap. For example,

$$\Phi_1(f(p))) = W_1^\top f(p) + b_1 \tag{2}$$

, where $W_1^\top$ and b_1 denote the weights and biases in the domain adaptation layer, respectively. And the matching metric between sketch and photo branch share the same linear transform due to the Siamese branch setting.

By training this unified model of both modalities, each of the FGIR tasks will benefit via learning a shared latent representation between the two tasks. At the inference stage, we can construct either sketch-photo, or text-photo ranking/retrieval according to the sketch-photo and text-photo ranking score, denoted as $R_s(s,p)$ and $R_s(t,p)$, respectively:

$$R_s\left(s,p\right) = -D\left(\Phi_1(f(s)),\Phi_1(f(p))\right) \tag{3}$$

$$R_s\left(t,p\right) = -D\left(g\left(t\right),\Phi_2(f(p))\right) \tag{4}$$

Figure 3: Example of the shoe multi-modal dataset.

4 Fine-grained Multi-modal Retrieval Dataset

We contribute a new dataset for multi-modal learning, especially for fine-grained cross-modal retrieval. Particularly, we collected a total of 1374 sketch-photo-text triplets for shoes. We first download shoe photos (in side view) and their corresponding descriptions from an online shopping website. To collect sketch-photo pairs, we follow [33] and ask volunteers to draw from mental recollections of the shoes photos: the sketcher was first shown a photo for 15 seconds; the picture was then withdrawn and he/she must draw the object from mental recollection. Some examples are given in Figure 3. [1]

We split shoe samples from each subcategory (boots, heels, sandals, slippers and so on) with the ratio 4:1, to form the train and test set. In total, we have 1112 sketches, text and photos to train our multi-modal deep neural network, and 262 instances for testing.

5 Experiments

5.1 Implementation details

Data preprocessing and augmentation We first pre-process photos into edge maps using EdgeBox [36]. We then do random crop and flip on both sketches and photos to augment the training data. Similar to other preprocessing strategies in text modality, we remove all stop words and symbols in the raw text description, as well as some rare words whose maximum word count are less than 5.

Network implementation We implement our multi-modal network in Tensorflow. Before fine-tuning on our dataset, we follow similar pre-training stages as detailed in [33]. More specially, we first pre-train the sketch and photo branches on TU-Berlin dataset [8] and extracted edges from ImageNet [33], respectively. For the text branch in our model, we first use gensim word2vec model [18] (pretrained on Google News dataset) on our training text description to pre-train the word embedding. We then fine-tune our model on our newly collected dataset with batch size set to 128. We use stochastic gradient descent (SGD) as the optimizer to train our multi-modal model, with a learning rate of 0.0001. Dropout is applied to both the FC layers and LSTM cells, with a keep rate of 0.5. We further put a l_2 regularizer with 0.0005 weight decay to reduce over-fitting.

Sampling strategy Our quadruplet sampler is inspired by the sampling strategy proposed in [10]. We first fix the ground-truth photo as the positive instance for both anchor sketch and anchor text. We then select 100 nearest neighbour photos in the VGG feature space [20] as hard negatives for each sketch and its corresponding text and ground-truth photo. As a result, a total of 111,200 quadruplets are generated for training prior to data augmentation. Such a

[1] The dataset will be made available at the SketchX website: http://sketchx.eecs.qmul.ac.uk/downloads/

sampling strategy largely mitigates the otherwise expensive human annotation employed by prior art [33].

5.2 Results

We compare our multi-modal fine-grained model with several baselines. First, we show the performance of our model is superior to baseline methods on the fine-grained top 1 and top 10 retrieval performance metrics. Then we present various ablation studies to cast insight on modality alignment, and how sketch-photo and text-photo can benefit each other.

Comparative results against baselines Three baselines are selected for comparison. The first is multi-view shallow CCA [7]. To obtain the multi-modal representations for this multi-view CCA framework, Sketch-A-Net [34] features (pool5 layer) are extracted for photo edges and sketches, respectively, while bag-of-words are applied to encode the text description. In the deep CCA baseline [26], one hidden fully-connected layer with 256 dimension transformed the same deep representation, and the CCA layer (32D) after that project the multi-modal embedding to the shared correlated latent space. The deep CCA model is learned via optimising the sketch-photo correlation and text-photo correlation, alternatively. Another baseline is a three branch (sketch, photo, and text branch) deep model, with two l_2 loss to match the embedding between sketch and photo, and between text and photo. Results in Table 1 demonstrate that our proposed method are clearly superior to the other baselines. It is also interesting to observe that both shallow CCA loss and deep CCA are not suitable for fine-grained retrieval, when compared to the l_2 loss and our unified quadruplet loss. We suspect this is because the feature space is already highly correlated, due to the many highly similar images in our gallery dataset.

Table 1: Comparative results against baselines on fine-grained SBIR and TBIR performance.

Model	sketch → photo		text → photo	
	Top 1	Top 10	Top 1	Top 10
Multi-view CCA[7]	0.38%	4.20%	0.76%	4.58%
Deep CCA[26]	7.25%	11.83%	0.38%	4.96%
Deep model + L2 loss	33.97%	72.14%	1.53%	5.73%
Our full model	**50.38%**	**84.73%**	**12.60%**	**37.40%**

Benefit from each cross-modal learning Our multi-modal learning model can also be viewed as a multi-task learning model, which has proven useful in many computer vision problems. In multi-task learning, each task can regularise the others, thus reducing over-fitting and promoting generalisation. In the deep learning context, this means they both provide more data to help to train 'latent tasks' in the form of a shared representation.

In our multi-modal framework, one task is fine-grained sketch-to-photo retrieval, while the other is fine-grained text-to-photo retrieval. The shared latent task is mining both the high semantic-level information (with the help of text modality) and also the low-level of structure and texture information (with the help of sketch modality) from the photo modality. In the ablation study, we first split our multi-modal model to two single-task cross-modality learning models, *i.e.*, the fine-grained SBIR and fine-grained TBIR models. We also train our full model by jointly training the two retrieval tasks. The retrieval performance is evaluated on our multi-modal dataset, as shown in Table 2.

Table 2: Contribution and performance of component tasks.

Model	sketch → photo		text → photo	
	Top 1	Top 10	Top 1	Top 10
Sketch-photo model	49.24%	82.06%	–	–
Text-photo model	–	–	8.78%	33.97%
Our full model	**50.38%**	**84.73%**	**12.60%**	**37.40%**

Performance comparison on fine-grained SBIR The multi-modal image retrieval task can be separated to fine-grained SBIR and TBIR tasks, and the fine-grained SBIR performance can also be evaluated with the sketch-photo subset of our multi-modal dataset. Here we compared our sketch-photo model with two most recent state-of-the-arts: triplet Sketch-A-Net [33] and triplet GoogleNet [19]. The results in Table 3 shows that both our sketch-photo model and Triplet Sketch-A-Net model works well, while ours can achieve the best top 1 and top 10 accuracy. Triplet GoogleNet can achieve similar performance compared to the triplet Sketch-A-Net model, but may suffer the over-fitting problem with more parameters.

Table 3: Performance comparison on fine-grained SBIR.

Model	Top 1	Top 10
Triplet Sketch-A-Net[33]	46.56%	82.82%
Triplet GoogleNet[19]	45.42%	79.77%
Our sketch-photo model	49.24%	82.06%
Our full model	**50.38%**	**84.73%**

Photo-text embedding alignment performance We evaluate against captioning approach to FG-TBIR by applying the CNN-RNN architecture as detailed in [23] to generate descriptions for our gallery images, and then perform text-to-text search. Another baseline model here is the deep CCA model, but with only two modalities as oppose to all three used in earlier experiments. From the results shown in Table 4, our text-photo model and caption model can achieve similar retrieval performance and are better than the deep CCA method. However our full multi-modal framework achieves the best performance.

Table 4: Photo-text embedding alignment performance with different methods.

Model	Top 1	Top 10
Multi-view CCA[7]	0.38%	5.34%
Deep CCA[26]	3.05%	18.70%
Photo caption model[23]	7.60%	24.40%
Our text-photo model	8.78%	33.97%
Our full model	**12.60%**	**37.40%**

Qualitative Results With our multi-modal retrieval model, we can apply the trained model to both sketch-to-photo retrieval and text-to-photo retrieval. Our model shows good performance on fine-grained SBIR, and the visual results of our proposed multi-modal framework is given in Fig. 4, where the ground-truth photo is highlighted using a green bounding box.

For text-to-photo retrieval, we test the model by giving the text description in the testing dataset, and then retrieve photos from the image gallery. For instance, if query text is given, the most similar photos retrieved are shown as Fig. 5.

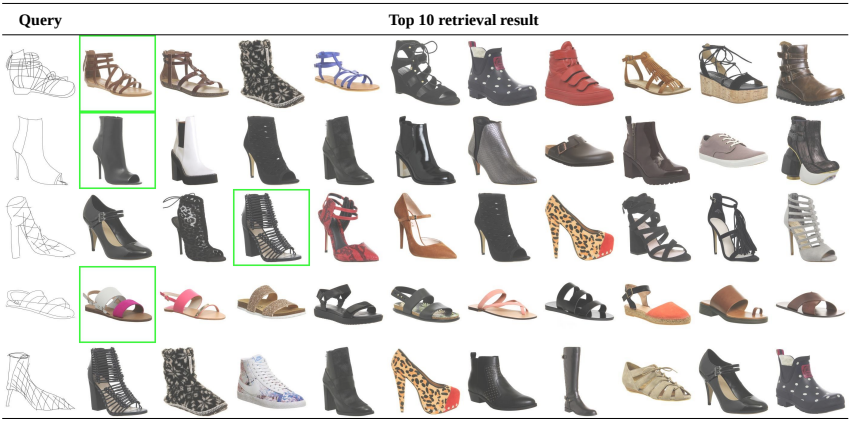

Figure 4: Example of fine-grained sketch-based image retrieval.

Figure 5: Example of fine-grained text-based image retrieval.

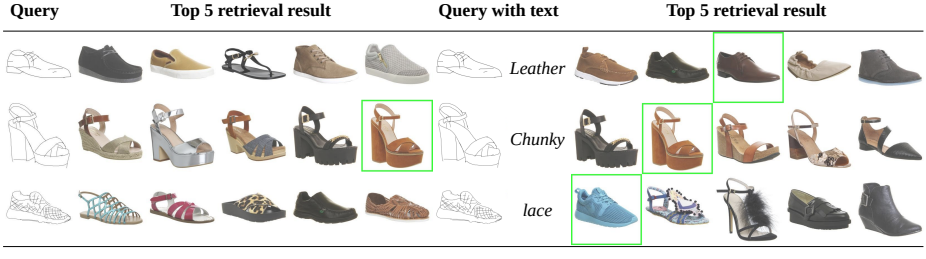

Figure 6: Example of fine-grained image retrieval with both sketch and text query.

Further insights on multi-modal query retrieval An unique characteristic of our model, compared with all previous fine-grained retrieval methods, is that it simultaneously embed all three modalities. By simply averaging sketch-photo and text-photo ranking scores ($R_s(s,p)$ and $R_s(t,p)$, respectively), we are able to use multi-modal query to conduct retrieval, *i.e.*, instead of using sketch alone, we could feed in sketch and text under one query to make retrieval even more fine-grained and comprehensive. For example, as Fig. 6 shows, when given a sketch query to the trained model, the network is able to retrieve structurally similar shoes. Yet it was not until text is added that the model could return true matches. This is because sketch can not convey features like material and fine texture, which are however straightforward to describe in text. Note that because we convert photos into edge maps on the photo branch following [33], our network is unable to encode colour information. We plan to alleviate this problem in the future by directly feeding colour photos into the photo branch as per [19]. An quantitative evaluation of sketch+text multi-modal retrieval, along with more qualitative examples, can be found in Supplementary Material.

6 Conclusion

In this paper, we proposed a multi-modal fine-grained retrieval framework, and also contribute a multi-modal FGIR dataset, where each sample has a photo, corresponding sketch and text. We investigate fine-grained SBIR and TBIR, showing that sketch is more powerful in isolation, but with a shared representation, both can be improved. Experiment results show that with the proposed multi-modal framework, our model can achieve a good retrieval result both on fine-grained sketch-to-photo and text-to-photo retrieval. Moreover, we offer insights on multi-modal query where sketch and text can be combined at testing time to obtain the most accurate results.

References

[1] Michael M Bronstein, Alexander M Bronstein, Fabrice Michel, and Nikos Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *CVPR*, 2015.

[2] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *TVCG*, 2011.

[3] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? In *SIGGRAPH*, 2012.

[4] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *CVPR*, 2015.

[5] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? Dataset and methods for multilingual image question. In *NIPS*, 2015.

[6] Mengyue Geng, Yaowei Wang, Tao Xiang, and Yonghong Tian. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*, 2016.

[7] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 2014.

[8] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *ICASSP*, 2013.

[9] Rui Hu, Mark Barnard, and John Collomosse. Gradient field descriptor for sketch based retrieval and localization. In *ICIP*, 2010.

[10] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *ICCV*, 2015.

[11] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[12] Ke Li, Kaiyue Pang, Yi-Zhe Song, Timothy Hospedales, Honggang Zhang, and Yichuan Hu. Fine-grained sketch-based image retrieval: the role of part-aware attributes. In *WACV*, 2016.

[13] Yi Li, Timothy M Hospedales, Yi-Zhe Song, and Shaogang Gong. Fine-grained sketch-based image retrieval by matching deformable part models. In *BMVC*, 2014.

[14] Yi Li, Yi-Zhe Song, Timothy M Hospedales, and Shaogang Gong. Free-hand sketch synthesis with deformable stroke models. *IJCV*, 2017.

[15] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. *arXiv preprint arXiv:1703.05605*, 2017.

[16] Xinyan Lu, Fei Wu, Siliang Tang, Zhongfei Zhang, Xiaofei He, and Yueting Zhuang. A low rank structural large margin method for cross-modal ranking. In *SIGGRAPH*, 2013.

[17] Kaiyue Pang, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. Cross-domain generative learning for fine-grained sketch-based image retrieval. In *BMVC*, 2017.

[18] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010.

[19] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: Learning to retrieve badly drawn bunnies. In *SIGGRAPH*, 2016.

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

[21] Jifei Song, Yi-Zhe Song, Tao Xiang, Timothy Hospedales, and Xiang Ruan. Deep multi-task attribute-driven ranking for fine-grained sketch-based image retrieval. In *BMVC*, 2016.

[22] Jifei Song, Yu Qian, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *ICCV*, 2017.

[23] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[24] Changhu Wang, Zhiwei Li, and Lei Zhang. Mindfinder: image search by interactive sketching and tagging. In *Proceedings of the 19th international conference on World wide web*, pages 1309–1312. ACM, 2010.

[25] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. In *CoRR*, 2016.

[26] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *ICML*, 2015.

[27] Xinggang Wang, Xiong Duan, and Xiang Bai. Deep sketch feature for cross-domain image retrieval. *Neurocomputing*, 207:387–397, 2016.

[28] John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*, 2015.

[29] L. Xie, J. Wang, B. Zhang, and Q. Tian. Fine-grained image search. *IEEE Transactions on Multimedia*, 17(5):636–647, 2015.

[30] Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. In *CVPR*, 2015.

[31] Licheng Yu, Eunbyung Park, Alexander C Berg, and Tamara L Berg. Visual madlibs: Fill in the blank description generation and question answering. In *ICCV*, 2015.

[32] Qian Yu, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Sketch-a-net that beats humans. In *BMVC*, 2015.

[33] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen Change Loy. Sketch me that shoe. In *CVPR*, 2016.

[34] Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Sketch-a-net: A deep neural network that beats humans. In *IJCV*, 2017.

[35] Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao. Sketchnet: Sketch classification with web images. In *CVPR*, 2016.

[36] Larry Zitnick and Piotr Dollar. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014.