Cross-Modal Learning for Sketch Visual Understanding

Jifei Song

A thesis submitted to the Queen Mary University of London for the degree of Doctor of Philosophy

School of EECS, QMUL, UK

June 2019

Abstract

As touching devices have rapidly proliferated, sketch has gained much popularity as an alternative input to text descriptions and speeches. This is due to the fact that sketch has the advantage of being informative and convenient, which have stimulated sketch-related research in areas such as sketch recognition, sketch segmentation, sketch-based image retrieval, and photo-to-sketch synthesis. Though these field has been well touched, existing sketch works still suffer from aligning the sketch and photo domains, resulting in unsatisfactory quality for both fine-grained retrieval and synthesis between sketch and photo modalities. To address these problems, in this thesis, we proposed a series novel works on free-hand sketch related tasks and throw out helpful insights to help future research.

Sketch conveys fine-grained information, making fine-grained sketch-based image retrieval one of the most important topics for sketch research. The basic solution for this task is learning to exploit the informativeness of sketches and link it to other modalities. Apart from the informativeness of sketches, semantic information is also important to understanding sketch modality and link it with other related modalities. In this thesis, we indicate that semantic information can effectively fill the domain gap between sketch and photo modalities as a bridge. Based on this observation, we proposed an attribute-aware deep framework to exploit attribute information to aid fine-grained SBIR. Text descriptions are considered as another semantic alternative to attributes, and at the same time, with the advantage of more flexible and natural, which are exploited in our proposed deep multi-task framework. The experimental study has shown that the semantic attribute information can improve the fine-grained SBIR performance in a large margin.

Sketch also has its unique feature like containing temporal information. In sketch synthesis task, the understandings from both semantic meanings behind sketches and sketching process are required. The semantic meaning of sketches has been well explored in the sketch recognition, and sketch retrieval challenges. However, the sketching process has somehow been ignored, even though the sketching process is also very important for us to understand the sketch modality, especially considering the unique temporal characteristics of sketches. in this thesis, we proposed the first deep photo-to-sketch synthesis framework, which has provided good performance on sketch synthesis task, as shown in the experiment section.

Generalisability is an important criterion to judge whether the existing methods are able to be applied to the real world scenario, especially considering the difficulties and costly expense of collecting sketches and pairwise annotation. We thus proposed a generalised fine-grained SBIR framework. In detail, we follow the meta-learning strategy, and train a hyper-network to generate instance-level classification weights for the latter matching network. The effectiveness of the proposed method has been validated by the extensive experimental results.

Declaration

I, Jifei Song, declare that this thesis titled, 'Cross-Modal Learning for Sketch Visual Understanding' has been composed by myself and the work presented in it are my own. I confirm that it has not been submitted, either in the same or different form, to this or any other university for a degree. All the quotation are distinguished by quotation marks, and all sources of information have been clearly acknowledged.

Some parts of the work have previously been published (or submitted) as:

- **Chapter 3**
  - J. Song, Y. Song, T. Xiang, T. Hospedales and X. Ruan, "Deep Multi-task Attribute-based Ranking for Fine-grained Sketch-based Image Retrieval", *BMVC*, 2016.
- **Chapter 4**
  - J. Song, Y. Song, T. Xiang, and T. Hospedales, "Fine-Grained Image Retrieval: the Text/Sketch Input Dilemma", *BMVC*, 2017.
- **Chapter 5**
  - J. Song, K. Pang, Y. Song, T. Xiang, and T. Hospedales, "Learning to Sketch with Shortcut Cycle Consistency", *CVPR*, 2018.
- **Chapter 6**
  - J. Song, Y. Yang, Y. Song, T. Xiang, and T. Hospedales, "Generalizable Person Re-identification by Domain-Invariant Mapping Network", *CVPR*, 2019.

Acknowledgments

I would like to first convey my grateful thanks to my supervisor Dr. Yi-Zhe Song. He offers me the precious chance to further my study at Queen Mary University of London. He leads me to this interesting topic, and gives me lots of valuable guidance during my research. Besides, his encouragement always gives me strength when I suffered from obstacles. I do wish he has a bright future in University of Surrey. Secondly, I would like to give great thanks to my co-supervisor Dr. Tao Xiang. He gives me a lot of technical suggestions for my research work. I indeed learned a lot from him on how to breakdown a challenging problem and solve them in a scientific way. I also have a close work with Dr. Timothy Hospedales before he moved to University of Edinburgh. I appreciate the theoretical help from him in the cooperation. At last, I am also grateful to Dr. Ioannis Patras for being my independent assessor to give me inspiring suggestions at my different stages of PhD study, and Dr. Yongxin Yang for the valuable help on recent researches in my last year.

At the same time, Tim Kay, and Harry Krikelis from EECS IT Service team, and Edward Hoskins from EECS health and Safety team, and kindly give me plenty of support when I work in EECS. I will remember the help from you, especially all the "urgent support" during the deadline days.

Both the academic time and spare time in London is cheerful to me. I enjoy the time with my dear colleagues and housemates. Special thanks to Yi Li, Xun Xu, Xiatian Zhu, Anran Qi, Feng Liu, Qi Dong, Conghui Hu, Kaiyue Pang, Ke Li, Da Li, Qian Yu, Tianyuan Yu, Jiabo Huang and Ying Zhang for the happiness when we gathered together.

Finally and sincerely, my heartiest thanks to my parents, for all the love and education

throughout the time when you bringing me up, and sincere thanks to my girlfriend, for the sharing of happiness and sorrow, and mutual help and belief for a long time.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

SBIR	Sketch-Based Image Retrieval
FG-SBIR	Fine-Grained Sketch-Based Image Retrieval
DNN	Deep Neural Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
GAN	Generative Adversarial Networks
3D	Three Dimensional
IID	Independent and Identically Distributed
GMM	Gaussian Mixture Model
FC	Fully Connected
LSTM	Long Short Term Memory
GRU	Gated Recurrent Unit
DIMN	Domain-Invariant Mapping Network
KL	Kullback Leibler
DPM	Deformable Part Models

Chapter 1

Introduction

Sketch research has attracted great interest for a long time, especially since the flourishing of touch-screen devices in recent decades [3]. Sketch research encompasses popular topics including sketch recognition [1, 7–10], sketch-based image retrieval [11, 11–13], fine-grained sketch-based image retrieval [3, 14–16], sketch synthesis [4, 17, 18], and others [19–21]. All these sketch-related tasks require a thorough understanding of sketch modality. For example, in a sketch-recognition task, the model needs to understand the semantic meaning of given sketches [1]. In fine-grained sketch-based image retrieval, an even more advanced, fine-grained understanding is required in addition to and understanding of the semantic meaning behind sketches [3]. 'Sketch' is mostly understood to has the meaning of a static modality in both sketch-recognition tasks and sketch-based image-retrieval tasks. However, in a sketch-synthesis task, the sketching process itself, as a kind of human drawing process, must also be understood and modelled. Knowing how the sketching process works can actually help us understand one of the unique attributes of sketches: the temporal, while semantic and fine-grained understandings describe the informative characteristics of sketches. These unique features of sketches make sketching a preferred modality for sketch-related research, assisted by existing sketch datasets

Dataset	Statistics	Sketch		Photo	
		Train	Test	Train	Test
TU-Berlin	# Categories	250	250	–	–
	# Instances/Category	54	26	–	–
QuickDraw	# Categories	345	345	–	–
	# Instances/Category	70,000	2,500	–	–
Skechy	# Categories	125	125	125	125
	# Instances	11,250	11,250	1,250	1,250
	# Images	65,064	6,312	11,250	1,250
QMUL-Shoe	# Instances	304	115	304	115
	# Images	304	115	304	115
QMUL-Chair	# Instances	200	97	200	97
	# Images	299	97	299	97
QMUL-ShoeV2	# Instances	1,800	200	1,800	200
	# Images	6,051	679	1,800	200
QMUL-ChairV2	# Instances	300	100	300	100
	# Images	951	324	300	100

Table 1-A: Dataset statistics for the popular sketch datasets.

like TU-Berlin, QuickDraw, QMUL-Shoe/Chair and so on (See Tab. 1-A for the statistics of different sketch datasets). However, drawing sketches and annotating the pair information is extremely expensive; developing a generalised model for sketch-related tasks is important because we will be able to get an acceptable performance in upcoming categories directly, thus saving expensive sketch collection and annotation for new categories.

The unique attributes of sketches have inspired many sketch-related research studies with promising application value. For example, studies in the semantic information and fine-grained details of sketches have enabled fine-grained sketch-based image retrieval. Studying the temporal aspects of sketches has helped us understand the sketching process and provide more human-like solutions for sketch synthesis tasks. Due to the great expense of collecting sketches and pairwise annotations, we decided to investigate the general abilities of sketch models, in order to avoid the expensive annotation of novel categories. Thus the contents of this chapter will be organised in three sections according to the three previously mentioned sketch-related tasks: fine-grained sketch-based image retrieval, sketch synthesis, and the general learning of fine-grained sketch-based image

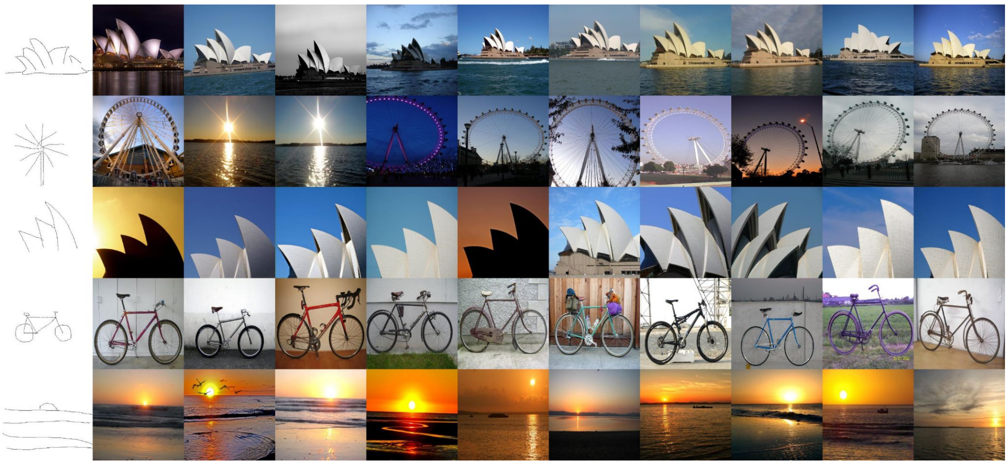


Figure 1.1: An example of sketch-based image retrieval.

retrieval, which together introduce the main applications of sketch research.

1.1 Fine-grained Sketch-based Image Retrieval

Sketch is informative. This advantage makes sketch modality one of the most appropriate modalities for retrieval tasks, especially in situations where photos are not readily available. This specific retrieval task is called sketch-based image retrieval (SBIR) and it presents one of the most fundamental problems in sketch research: how to retrieve the photos in a gallery that match with a given sketch. Previous SBIR projects [11] typically used hand-crafted features and assembled them using a bag-of-words framework. Recent SBIR projects use a deep learning framework to achieve better performance, as the learned features are usually more powerful than the hand-crafted features [22]. One of the drawbacks of traditional sketch-based image retrieval is that it only treats a sketch as a semantic modality similar to a text-based description, neglecting the fine-grained details embedded in the sketches. Figure 1.1 presents some typical examples of sketch-based image retrieval.

Fine-grained SBIR tasks require the model to find the exact photo required by a given query sketch, rather than returning any given photo from the same category. This demonstrates the unique advantages of informativeness, making sketch modality more effective than other modalities such as text in instance-level image retrieval. Yi *et al.* first tried to encode the fine-grained features of sketches via deformable part-based models, while Qian *et al.* later developed the deep version of the fine-grained SBIR model with satisfying results.

The problem with most existing fine-grained SBIR methods is that they ignore the semantic connection between sketch and photo modality, which constrains the model's understanding to a feature level rather than a semantic level, thus impeding the model's ability to achieve a satisfactory result. We find that attributes can play this semantic role, helping the fine-grained SBIR model bridge the domain gap between sketch and photo at the semantic level. We then proposed a deep attribute-driven multi-task framework [23] to exploit the attribute information and improve the fine-grained SBIR performance. Finally, we compared the sketch modality to the text modality, and discovered that text can also be used to help the fine-grained SBIR task via the additional semantic information using our multi-modal framework [24].

1.2 Sketch Synthesis

Temporal information is another valuable attribute of sketches clearly reflected in the sketching process. Understanding the sketching process is the first step in creating advanced sketch tasks like photo-to-sketch synthesis. Sketch synthesis is another challenging application of sketch research: creating the ability to render photos into human-like sketch drawings.

Similar to other vision tasks, before overwhelmed by deep learning methods, most sketch synthesis methods followed the process of extracting the edges from photos, then either deforming the edges into sketch-like strokes or replacing edges with matching strokes [17, 25]. The generative adversarial network (GAN) [26] has achieved amazing results in image style transfer tasks, and outperforms most existing non-deep methods.

GAN was also applied on photo-to-sketch translation in [27], though they mainly worked on facial sketches, which have less deformation and fewer abstraction gaps than free-hand sketches.

However, treating sketch synthesis as one kind of image translation task with 'sketch style' as the target style will cause a host of problems: first, the temporal/sequential drawing process in sketching is ignored in these GAN-based methods [26, 27]. Second, existing image translation struggles with free-hand sketch synthesis, as it requires that the input photo needs to be aligned with the target sketch, which is not suitable for free-hand sketches as there is an abstraction gap between free-hand sketches and photos. Therefore, we proposed a CNN-RNN framework [28] to model both static image content and the temporal drawing process, and it has surpassed existing methods.

1.3 Generalised Learning for Fine-grained SBIR

Though sketches have a number of valuable attributes including fine-grained information, collecting sketches and annotating the pair information is extremely time-consuming and expensive. The expense of collecting sketches has motivated research into generalised learning for fine-grained SBIR and similar tasks. Generalisation problems arise when applying trained fine-grained SBIR models to practical scenarios where the input categories might be unseen before. Thus the concept of generalised learning for fine-grained SBIR is proposed to strengthen the generalisation ability of fine-grained SBIR models.

The uncertainty of new testing categories in the fine-grained SBIR system will make existing fine-grained SBIR models suffer, as they will more or less overfit to known categories in the training set. We observed that the parameters of existing fine-grained SBIR deep models were highly dependent on training categories. Therefore, we consider learning to generate the parameters of the matching model based on individual instances. Specifically, we propose a generalised fine-grained SBIR model, where a HyperNetwork is introduced to synthesise the parameters for the matching module based on current instance information, independent of training categories. Experimental results have shown that the proposed method can achieve satisfying performance in generalisation and has outperformed baseline methods.

1.4 Contributions

The contributions of this thesis are:

1)For Fine-grained SBIR:

We propose an attribute-driven fine-grained SBIR framework that exploits attributes as semantic clues to improve the fine-grained SBIR performance. Experiments show that attributes can help the fine-grained SBIR task on both “shoe” and “chair” datasets.

We also investigate the fine-grained model’s ability to compare between sketches and photos and implement a multi-modal framework to fulfil the multi-modal retrieval task between sketch, text and photo modalities. We discover that text can also be used as a semantic bridge between sketch and photo modalities. We collect a multi-modal dataset with 2,000 sketch-text-photo tuples to aid the multi-modal retrieval research. Experiment results show that the proposed multi-modal multi-task framework can effectively achieve satisfying performance for the multi-modal retrieval task.

2)For Sketch Synthesis:

We propose a CNN-RNN architecture to capture semantic information in given photos and draw sketches segment by segment with the supervision of ground-truth sketches. Both qualitative and quantitative results show that the proposed framework can outperform state-of-the-art methods. We further show that our methods can also be employed to generate synthetic data as the pre-training data for the fine-grained SBIR task.

3)In the Generalised Fine-grained SBIR:

To address the generalisation problem in fine-grained SBIR, we proposed a generalised learning framework for fine-grained SBIR. We took a meta-learning strategy to help our framework learn a generalised metric for fine-grained SBIR. During the inference stage, the model can generate classifiers for new classes, which will then be used to predict the matching relations among target data. The proposed method can achieve better generalised performance than comparable methods.

1.5 Outline of the Thesis

From here, the thesis is organised as follows:

Chapter 2 presents a thorough literature review of the main areas of sketch research, *e.g.*, sketch recognition, fine-grained SBIR, sketch synthesis and the generalised learning for sketch-related tasks like fine-grained SBIR.

Chapter 3 proposes an attribute-driven multi-task framework for fine-grained SBIR, with the attribute functioning as a semantic bridge to narrow the domain gap between sketch and photo.

Chapter 4 covers a comparison among sketch, text and photo modalities, and introduce a framework for multi-modal retrieval.

Ch.	Method	Task	Contribution	Input	Output
III	FG-SBIR with Attri.	FG-SBIR	Consider semantic understanding from attributes; Better retrieval performance	Sketch Photo	Distance
IV	FG-SBIR with Text	FGIR	Accept more flexible inputs like text; Improve retrieval performance	Sketch Photo	Distance
V	Neural Sketcher	Sketch Synthesis	First deep approach for sketch-to-photo synthesis	Sketch Photo	Generated Sketch offsets
VI	DIMN	Generalised FG-SBIR	Achieve more generalised performance for FG-SBIR	Sketch Photo	Distance

Table 1-B: Contributions of each proposed method.

Chapter 5 addresses a novel design of a stroke-level photo-to-sketch framework, which can simulate the sketch-drawing process of a human.

Chapter 6 describes a generalised learning framework for fine-grained SBIR tasks, and evaluates the generalisation of the proposed method and its alternatives.

Chapter 7 presents the conclusion and some thoughts for future research.

A tabular summary of the contribution coming from each chapter is presented in Table. 1-B.

Chapter 2

Literature Review

This literature review presents a summary of existing methods of sketch applications such as fine-grained sketch-based image retrieval, sketch synthesis, and generalised learning for fine-grained sketch-based image retrieval. To provide a better understanding of these fields, the literature review also covers related sketch research topics such as sketch recognition, sketch-based image retrieval, domain generalisation and meta-learning. We review related works based on their topics in the following order: the state of sketch, fine-grained sketch-based image retrieval, sketch synthesis, generalised learning for fine-grained sketch-based image retrieval, review on sketch benchmarks.

2.1 The State of Sketch

Sketch has been used to record events since historical time. Since then, sketch has been widely applied in many format of human drawings, like free-hand sketches, professional sketches, and so on. Admittedly, professional sketches have significant usage for forensic science such as face synthesis [29] and recognition [30]. However, compared to free-hand sketches, professional sketches are out of reach for the common users without professional art training. In contrast, free-hand sketch is a more user-friendly human drawings, thus we mainly focus on researching with free-hand sketches in this thesis, and if not mentioned specifically in the following content, sketch in this thesis represents for free-hand sketch.

Sketch has shown many unique features beneficial for vision research. For example, sketch is first informative, containing semantic information which can be understood by sketch recognition techniques [1, 31]. Beyond semantic information, sketch also conveys fine-grained detailed information, facilitating the fine-grained sketch-based image retrieval [3, 14–16]. Such fine-grained details inherited in sketch modality make sketch a strong competitors for modalities used for object retrieval [3, 14–16, 32, 33] like text and photo.

Different with other informative modalities like photo, sketch also brings feature of temporal information. When human make sketch drawings, the stroke order, and also the coordinates flow in each stroke reflect the temporal feature inside this modality, while in comparison, photo is captured by image sensing from charge coupled device (CCD) cameras immediately, without any temporal information recorded other than the static pixel information. Such advantages of encoding temporal orders open the research on sketch generation [2] and sketch synthesis [17]. Moreover, sketching process also involves human's analysis and visual understanding for the given object. Thus researches on human sketch drawing to some extent open the gate for human visual understandings. The last unique attribute of sketches is the abstractness, removing the redundancy of normal drawings while keeping the essential components. Related researches are like sketch abstraction [21], sketch perceptual grouping [19, 20], and so on. In the following sections, we will give the literature review on the main sketch research areas driven by the aforementioned sketch unique attributes, such as sketch recognition, sketch-based image retrieval, fine-grained sketch-based image retrieval, sketch synthesis and so on.

2.2 Sketch Recognition

As a fundamental topic of sketch research, sketch recognition considers how to recognise a given sketch, most practically how to recognise free-hand sketches [1]. Sketch recognition also provides a useful understanding of closely correlated sketch tasks such as sketch-based image retrieval and sketch synthesis. A typical example of a sketch recognition task is shown in Fig. 2.2, with a query sketch and prediction results returned from the live demo we developed, which can be accessed at `https://sketchx1.eecs.qmul.ac.uk/`.

As a field of research, sketch recognition lacked a large-scale dataset, meaning earlier sketch recognition methods were primitive, and mainly worked on symbols and curves [7, 8]. The introduction of the TU-Berlin dataset has greatly improved research [1] with its carefully curated collection of free-hand sketches of both good quality and quantity. The TU-Berlin dataset is a large-scale object-level sketch dataset, which includes 20,000 sketches spanning 250 categories. Alongside the dataset, this work proposes a typical solution [1], which extracts orientation-specific hand-crafted features and applies a support vector machine (SVM) to classify the bag-of-features representation. An improved version in [9] further considers the assembly of local features using a star graph, and achieves better recognition results. Fisher vectors are also successfully applied to sketch recognition in one of the recent works [10]. Later on, in 2017, Google collected an even larger sketch dataset [2], with many more categories (345) and sketches (50 million), but the drawing quality is inferior to those in TU-Berlin. Besides supporting research into sketch recognition tasks, the temporal order recorded in QuickDraw datasets allow sketch synthesis methods [5] to interpret temporal information and model sketching processes more effectively. A comparison between representative sketches in TU-Berlin datasets and those in QuickDraw datasets are shown in Fig. 2.1, where we can see that TU-Berlin has a much better quality of data than the QuickDraw datasets in the given categories like ‘airplane’, ‘clock’, ‘elephant’ and ‘television’, while QuickDraw provides the largest-scale existing collection of sketches.

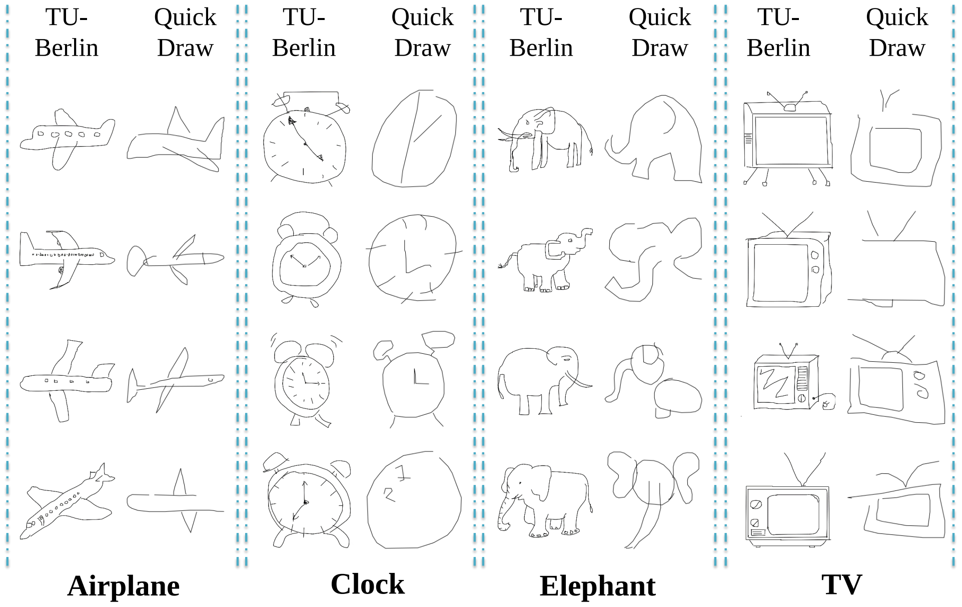
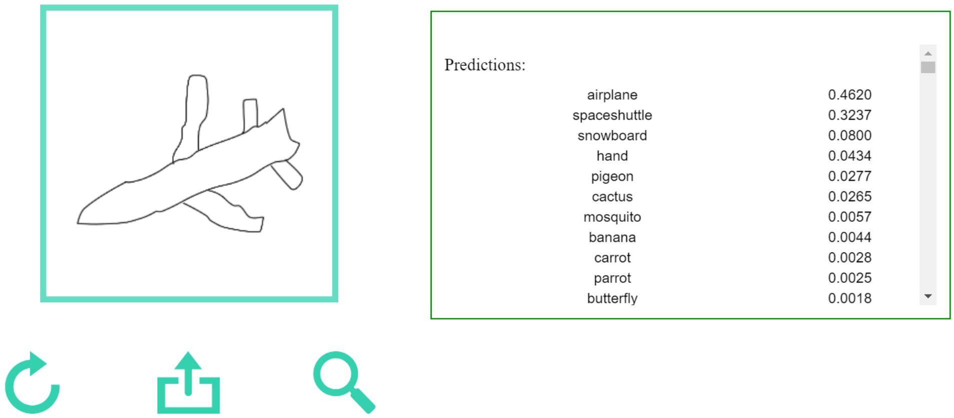


Figure 2.1: Comparison of representative sketches between the TU-Berlin dataset [1] and the QuickDraw dataset [2].

Deep learning has demonstrated superior performance when applied to sketch recognition tasks. Among all the deep learning-based sketch recognition methods, Sketch-a-Net [31] is the first deep network specifically designed for sketch recognition, which achieves the state-of-the-art performance and beats human as well. Wang *et al.* [34] applied PointNet [35] to sketch recognition, which can achieve comparable results to [31], but it engages far fewer parameters. In [36], a recurrent neural network (RNN) is used to temporally model the sketch feature based on a convolutional neural network (CNN); this achieved better performance than [31], but the drawback of this method is its huge time consumption. Another recent work [37] also incorporated RNN and CNN to achieve a satisfying performance, and at the same time embedded a hashing module to ensure scalable sketch recognition. In our work in fine-grained SBIR, we also use Sketch-a-Net as the backbone to encode the sketches, as it has been shown to develop a good understanding of free-hand sketches.



Predictions:

airplane	0.4620
spaceshuttle	0.3237
snowboard	0.0800
hand	0.0434
pigeon	0.0277
cactus	0.0265
mosquito	0.0057
banana	0.0044
carrot	0.0028
parrot	0.0025
butterfly	0.0018

Figure 2.2: An example of sketch recognition.

2.3 Sketch-based Image Retrieval

Sketch-based image retrieval is the ability to retrieve photos using sketches as input. Compared with traditional text-based image retrieval (TBIR) [38–40] and content-based image retrieval (CBIR) [41, 42], SBIR provides a more intuitive and convenient method to users. For example, it is easy to sketch the object in our mind, but harder to output with an identical image or a detailed description of it. Owing to its advantages, SBIR is very promising in commercial applications. The illustration of comparison among SBIR, TBIR and CBIR is shown in Fig. 2.3.

Most existing SBIR methods exploit the unique properties of sketches and find a bridge to cross the domain gap between photos and sketches. For example, inspired by the histogram of oriented gradients (HoG) descriptor, Hu *et al*. proposed a gradient field HoG (GF-HoG), which is specifically designed for sketch-based image retrieval tasks [11]. Later on, Hu applied GF-HoG feature into a bag-of-regions (BoR) representation to finish an SBIR process [12] and show improvements compared to the classical bag-of-words framework [11]. A newly designed descriptor called soft-histogram of edge local orientations (S-HELO) can also achieve good performance in this task [13].

A deep framework was proposed in [43] to achieve both SBIR and sketch recognition tasks, which customise AlexNet [44] to an R-Net and an S-Net for photo and sketch

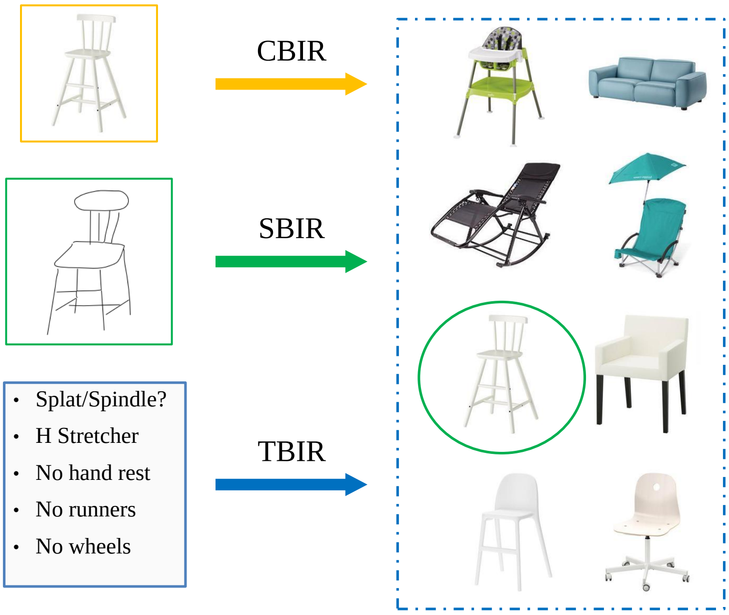


Figure 2.3: Comparison among SBIR, TBIR and CBIR.

branches, respectively. Bui *et al.* also tasked a CNN to learn a compact descriptor, with the help of triplet loss [45]. Hashing is also considered and embedded in CNN to accelerate the SBIR efficiency by learning concise hash code [22, 46, 47]. Recently, Radenovic *et al.* [48] proposed the EdgeMAC descriptor which is learned from a deep network and has shown good ability in SBIR.

However, traditional SBIR tasks focus on category-level retrieval. In other words, existing SBIR methods only care that the retrieved the photo comes from the same category as the query sketch. Fine-grained details are neglected, making it unable to retrieve the exact photo which is most similar to the query sketch and thus users might prefer TBIR to SBIR.

2.4 Fine-grained Sketch-based Image Retrieval

Fine-grained sketch-based image retrieval (FG-SBIR) aims to retrieve the target photo among very similar photos in the gallery according to the query sketch, which requires the designed model to capture the instance-level fine-grained details of both the sketches and the photos. This task is very challenging because of the difficulty of finding discriminative features and eliminating the domain gap between sketch and photo modalities. The retrieval examples returned from our demo in Fig. 2.4 intuitively present this interesting task and its promising application value.

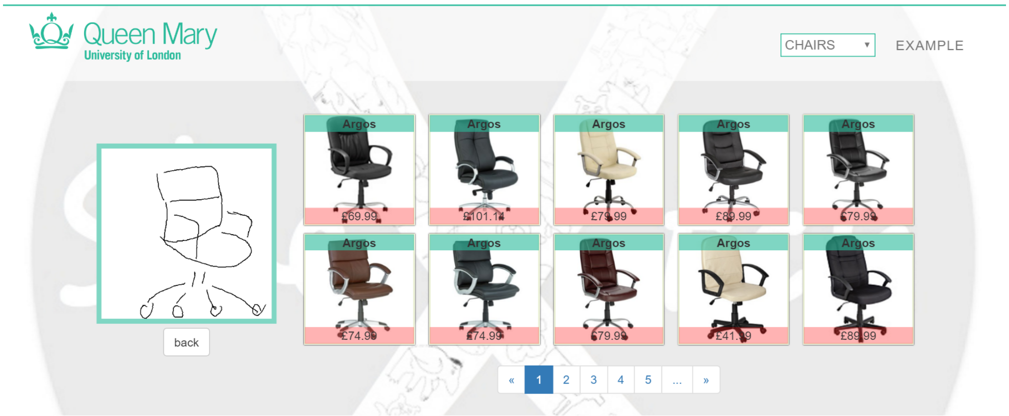


Figure 2.4: A retrieval example obtained through our FG-SBIR demo.

Yi *et al*. initially defined this problem in [15], and proposed an entire pipeline for FG-SBIR, which extracts HOG features from each sketch and photo, and encodes them through deformable part models (DPM) [49] with a graph-based part matching module followed to deal with pose changes. Inspired by this work, Ke *et al*. proposed a synergistic representation combining low-level, mid-level and high-level features, which has proved to be beneficial for fine-grained SBIR performance in experimental results. One recent work [50] summarised the popular cross-modal subspace learning methods and applied them to the fine-grained SBIR task.

Deep neural network (DNN) provides an end-to-end solution for fine-grained SBIR [3, 14], and has proven to be superior to shallow methods which are normally based on hand-crafted features [15, 16]. [3] formulates a triplet-ranking network to align sketch and photo modalities. It adopts a Siamese architecture, and utilises a triplet loss to learn a joint embedding space. The Sketchy network [14] uses a heterogeneous architecture instead and employs GoogleNet on each branch to learn modality-dependent feature representations. A triplet loss is used on the final fully-connected (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.

The problem in the existing deep learning based FG-SBIR method is that they neglect semantic information, thus limiting the performance of fine-grained SBIR models. In Chapter 3, we for the first time developed deep model-considering attribute-semantic information, forming a deep multi-task framework. Another problem unexplored in this field is how different the contribution of text and sketch modalities are to the fine-grained image retrieval problem, and whether these two modalities can help each other. Therefore, in Chapter 4, we introduce a multi-modal fine-grained image retrieval framework, adopting the former's Siamese architecture, and extending the triplet ranking loss to a quadruplet one in order to embed three modalities (sketch, text and photo).

2.5 Sketch Synthesis

Sketch synthesis refers to the ability to draw a sketch like a human according to a given photo. This is an extremely challenging task because photo and sketch domains differ significantly. Furthermore, human-drawn sketches exhibit various levels of sophistication and abstraction even when depicting the same object in a reference photo.

Previous sketch synthesis works follow a rendering pipeline by breaking down the translation from photo to sketch into various levels of abstraction to synthesise the sketch. One example of this kind of attempt is [25], which extracted edges from photos and replaces edges with strokes based on learned parameters reflecting the style and abstraction level. Another attempt learned a deformable stroke model (DSM) on human stroke data in a given category, and fit the learned DSM model to given photos and synthesised free-hand sketches, as presented in [17]. Recent advances in generative adversarial networks (GANs) make realistic image generation possible [4]. Zhang *et al.* take a classic image-to-image translation framework called pGAN to efficiently generate face sketches based on input photos. [18] develops a two-branch style-transfer network, which synthesises the final sketch by combining the content of a given photo and a certain target style.

The main problem for existing sketch synthesis methods is that they treat a sketch as one certain style of static image, thus neglecting the sequential information encoded in sketches. Therefore, sketches generated by most existing methods look like drawings with sketch-style edges, rather than real human sketches made by a human sketching process, which would be drawn stroke by stroke. We proposed to build ties between raster and vector sketch images through a CNN-RNN paradigm, as detailed in Chapter 5. In this work, sketches are modelled as sequential vectors and an RNN decoder is employed to draw sketches conditional on CNN encoder embedding.

2.6 Generalisation for FG-SBIR

Traditional FG-SBIR methods focus on learning fine-grained features in given categories. However, these methods will suffer when applied to real scenarios, where new unseen categories will factor into the designed FG-SBIR system. Existing deep FG-SBIR models that are directly applied to new categories without model updating are known to suffer from considerable performance degradation [3, 15, 51], thus leading to model overfitting and poor generalisation.

The domain generalisation (DG) [52, 53] problem is closely related to the mentioned task. DICA [53] proposed learning the domain-invariant features via a kernel-based optimisation. Recently, Motiian *et al.* extended a supervised domain adaptation network to domain generalisation by explicitly imposing a semantic alignment loss on every unpaired data set [54]. The idea of adversarial training for unseen domain data synthesis is exploited in CrossGrad [55], where pseudo-training scenarios are generated by perturbations in the direction of the gradients of the domain classifier and category classifier respectively. As an early attempt to apply meta-learning techniques to domain generalisation, MLDG [56] proposed to align meta-train and meta-test gradients by using the same training schedule, *i.e.*, task (re)sampling, where the idea is similar to the meta-learning model, MAML [57].

Meta-learning is also correlated to our task. Meta-learning [58] is a frequent topic in the machine learning community, and one of its well-received applications is few-shot learning (FSL). FSL aims to recognise novel visual categories from limited labelled examples in situations where conventional fine-tuning is unlikely to work due to overfitting. As one of the classic meta-learning methods, prototypical networks [59] proposed learning a prototype for each class, where the classification is based on computing the distances to those prototypes. Instead of using the prototype to generate the linear classifier, PPA [60] learned to derive classifier parameters from the average supporting activations.

In Chapter 6, we first consider the generalisation problem in fine-grained sketch-based image retrieval. Note that this problem is much more challenging than the conventional FG-SBIR problem, as target categories/instances are different from source ones, which means we have to deal with domain gap and disjoint label space simultaneously. We propose a generalised framework for FG-SBIR and show that our method is much more effective than a number of baseline methods, due to its unique end-to-end image-to-classifier learning.

2.7 Sketch Benchmarks

Sketch datasets have set the benchmarks for different sketch research fields, and drive the corresponding researches. A good sketch dataset should include common sketching styles as many as possible and also cover a large variety of categories/instances. The collection of sketch dataset is normally costly and also time consuming as many amateur volunteers need to be recruited to draw sketches within acceptable quality and carefully label them.

“TU-Berlin” dataset [1] is the first large-scale database in sketch recognition community. There are 20,000 sketches spanned 250 categories in this dataset. As one of the most popular benchmark in this field, TU-Berlin dataset has witnessed many outstanding deep and non-deep approaches in sketch recognition area [1, 9, 10, 31]. Later, a much larger database,“QuickDraw” [2] has been published by Google in 2017. QuickDraw dataset has provide much more (75,000 vs 80) sketches for each category, and also much more (345 vs 250) categories are considered. Apart from the contribution of a larger dataset in sketch recognition community, the vast number of vectorized sketch drawings are also helpful for exploiting the sequential information standalone [21] or alongside with the static visual information [37].

When go to sketch-based image retrieval, only dataset with pair information can be used to learn the matching between sketch and photo and also evaluate the learned matching function. Flickr15K [45] is widely used for traditional category-level sketch-based image retrieval in the beginning stage, with 330 sketches and 10,000 images across 33 categories, paired in a category level. However, the matching ability learned in the traditional sketch-based image retrieval task remains in a rather coarse level, as the fine-grained details has been largely omitted, and that’s why fine-grained sketch-based image retrieval has now become more and more popular in retrieval task. An initial contribution in fine-grained sketch-based image retrieval is the intra-category sketch retrieval benchmark proposed in [15], considering 14 categories with in total 1,120 sketches and 7,267 photos. A more well-defined benchmark is then proposed in [3], namely QMUL-Shoe and QMUL-Chair, where the sketches are matched in instance-level by asking volunteers to draw sketches according to the given object photos. Sketchy dataset is the latest and largest published dataset in this field, collecting 12,500 photos and 75,471 fine-grained paired sketches.

In the field of photo-to-sketch synthesis, the datasets need to be with fine-grained paired photos and sketches, as well as the sequential information of the sketch drawings, which is demanded by the nature of sketching process. Sketch synthesis is for now supported by the datasets like QMUL-ShoeV2 [61], QMUL-ChairV2 [61], and QuickDraw dataset [2]. Apart from these mentioned sketch related tasks, sketch has even wider usage for different research topics, with corresponding sketch datasets supporting these researches. For example, a scene sketch dataset is proposed in [62] to aid the research in scene-level sketch synthesis and retrieval. Moreover, SHREC'13 [63], SHREC'14 [64], and SHREC'16 [65] databases is helpful for sketch-based 3D retrieval. Specifically, sketches collected in SHREC'16 benchmark are 3D sketches, while the others are normal 2D sketches. Though there already exist many sketch databases, more and more sketch-related databases are still demanded with the high pacing progress of sketch-related researches. In this thesis, we also contributed a multi-modal fine-grained dataset, as detailed in Chapter 4.

2.8 Summary

2.8.1 Fine-grained Sketch-based Image Retrieval

Compared to the traditional SBIR task, fine-grained SBIR focuses on capturing fine-grained details in both sketches and photos, thus taking informative advantage of the sketches. We further investigate the important role of semantic information such as attributes and text descriptions to this problem.

For attributes, we propose a multi-task framework to leverage the semantic information in attributes to aid the fine-grained SBIR performance, as described in Chapter 3. For text description, we develop a multi-modal framework which can conduct both sketch-to-photo retrieval and text-to-photo retrieval. We also show that the two tasks can help each other in our unified framework.

2.8.2 Sketch Synthesis

Photo-to-free-hand sketch synthesis aims to mimic the sketching ability of a human, *i.e.*, to draw sketches stroke by stoke based on the analysis of a given object. We indicate that the sketch drawing process has somehow been neglected by most existing methods, where sketches are treated as static images rather than sequential vectors.

In Chapter 5, a deep photo-to-sketch synthesis model is proposed, which tries to understand given photos based on a CNN encoder while drawing sketches at stroke level with the help of RNN decoder. Auxiliary tasks like photo synthesis and sketch reconstruction are also considered to assist in the main task.

2.8.3 Generalisation for FG-SBIR

Generalisation for FG-SBIR considers the generalisation abilities of FG-SBIR models, which are not considered by existing FG-SBIR models despite their significant value for practical applications.

In Chapter 6, inspired by the idea of few-shot learning, we design a generalised learning framework for FG-SBIR using meta-learning strategies. The main idea of our framework is to classify matching relationships between the photo and sketch pairs, using generalised parameters learned across different meta-epochs.

Chapter 3

Fine-grained Sketch-based Image Retrieval with Attribute

With touch-screen devices becoming ever more ubiquitous, sketch possesses a great advantages as an intuitive and efficient mode of input compared to traditional alternatives such as text or speech. One of the most promising attributes is conveying fine-grained information. This unique promise of the sketch modality has motivated a major revival of interest in vision-based analysis of sketches, notably in sketch-based image retrieval (SBIR). Most existing SBIR methods operate at the category-level [11, 12, 66–70]: *i.e.*, retrieving photos of the object coming from the same category as the query sketch among photos from a set of categories. However this means that sketch as a query modality is in direct competition with text – the user typically can specify a category more clearly and easily using text, making SBIR a less appealing retrieval paradigm. In contrast, a more unique property of sketch is the ability to encode fine-grained visual details that would otherwise be hard to describe in text, especially considering there are some components need to be described with professional knowledge. This observation has led to the recent emergence of fine-grained SBIR [3, 14, 15].

Fine-grained sketch-based image retrieval (FG-SBIR) focuses on finding specific photos in the gallery that match as closely as possible the details encoded in the query sketch. Due to the drastic appearance differences across the sketch and photo domains, especially between free-hand sketches and photos, FG-SBIR is an extremely challenging task and very few attempts are reported. An earlier method in [15] extracts histogram of gradients (HOG) features from each sketch/photo and encodes them into deformable part models (DPM); a graph-based part matching is then followed to deal with the pose changes. In contrast to hand-engineering features, recently a deep learning approach is proposed [3], aiming to learn a higher-level feature representation with the right (in)variance properties across the sketch-photo domains jointly with the matching function. Specifically, a three-branch deep neural networks (DNN) is trained with a triplet ranking loss to match sketches to the corresponding photos. Optimising this objective function requires the network to re-represent the photo/sketch in an aligned embedding subspace to eliminate the domain gap while emphasising the fine-grained details. Similarly, a two-branch DNN is developed in [14] for instance-level SBIR, with Heterogeneous GoogleNet [71] used as the basic framework. While such DNNs outperform prior work based on hand-crafted features, their efficacy is limited by the lack of knowledge about the semantic properties shared by a matching sketch-photo pair. Moreover, in order to learn this triplet-ranking based DNN, fine-grained human annotations are required which are both costly and error-prone to generate: for any given query sketch, the number of ranking pairs of photos is quadratic of the number of photos; and many photos are visually too similar for even humans to differentiate reliably (as illustrated in Figure 3.2).

In this work, we wish to take advantage of a DNN's strength as a representation learner, but also combine this with semantic attribute learning, resulting in a deep multi-task attribute-based ranking model for FG-SBIR. In particular, we introduce a multi-task DNN model, where the main task is a retrieval task with triplet-ranking objective similar to [3], and attributes are detected and exploited in two auxiliary tasks. The first side-task is to predict the attributes of the input sketch and photo images. By optimising this task at training-time, we encourage the learned representation to more meaningfully encode the semantic properties of the photo/sketch. The second side-task is to perform retrieval ranking based on the attribute predictions themselves. At test-time, this means that the retrieval ordering is explicitly driven by semantic attribute-level similarity as well as the similarity of the internally learned representation. This novel deep multi-task attribute-based ranking network architecture has a number of advantages over existing methods: (1) The unique domain-invariant nature of visual attributes helps to bridge the cross-domain gap between photos and sketches. (2) By introducing multiple tasks in the network, the model generalises better and further can rely less on expensive human ranking annotation. Specifically, we show that the highly non-scalable step of triplet annotation required by the model in [3] can now be avoided and an automatic attribute-based strategy is developed instead to focus on the most informative 'hard' training samples for more efficient learning of the model.

It is worth noting that, although this is the first time a deep multi-task learning (MTL) approach is developed for FG-SBIR, similar approaches have been successfully applied to other vision problems to exploit the fact that different tasks can effectively regularise each other when solved simultaneously, thus allowing all tasks to generalise better to test data. For example, deep facial landmark detection task is improved when trained alongside facial attribute classification [72]: the representation necessary to support attribute prediction is also helpful for encoding the location of facial landmarks. In the video thumbnail selection problem, the image search task based on click-through is set as the side task while the main task is the deep visual-semantic embedding [73]. Another example is pedestrian attribute prediction improving the main task of pedestrian detection [74]. However, dealing with a cross-domain matching problem such as FG-SBIR has additional challenges which are addressed uniquely in this work by carefully designing learning tasks and strategies tailor-made for the fine-grained retrieval problem.

The contributions of this work are two-fold: (1) A novel deep MTL model is proposed to exploit two attribute-based auxiliary tasks for learning semantically meaningful and domain-invariant representation for FG-SBIR. (2) A new attribute-based triplet generation and sampling strategy is developed to boost the effectiveness of the deep MTL model. Extensive experiments are carried out on two benchmarks and the results demonstrate that the proposed model significantly outperforms the state-of-the-art while simultaneously requiring less costly annotation.

3.1 Methodology

3.1.1 Multi-task Fine-Grained SBIR Network

In this section we describe our multi-task deep neural network for fine-grained SBIR. The DNN architecture is illustrated in Figure 3.1. The proposed network is a three branch network. Each input tuple consists of three images corresponding to the query sketch (gone through the middle branch), positive photo image (top branch) and negative photo image (bottom branch) respectively. The positive photo has been annotated as more visually similar to the query than the negative photo. The learned deep model aims to enforce this ranking in the model output.

3.1.1.1 Network Structure

The network structure of the proposed model is shown in Figure 3.1. We take a similar encoder following the state-of-the-art FG-SBIR approach [3]. Basically, we have a triplet branch network, where there branches take the input of anchor sketch, positive photo and negative photo. We implement the Siamese netwwork for the proposed framwork, *i.e.*, the photo branches and the sketch branch share the parameters, which is designed to overcome the overfitting problem in expensive and limited sketch dataset [3]. The first part of the triplet branch network are shared by diffrent tasks, including main sketch-to-photo retrieval task and auxiliary attribute tasks, while the second part are designed separately for different tasks, which is a common network designing strategy in multi-task frameworks [72–74].

The architecture of the task-shared part consists of five convolution layers with max pooling, as well as a fully-connected (FC) layer, to learn a better representation of original data via feature maps. After these shared layers, different tasks evolve along separate branches: in the main task, one more FC layer with dropout and rectified linear unit (RELU) are added to represent the learned fine-grained feature vectors. Similarly, in the auxiliary task, another FC layer (with dropout and RELU) extracts fine-grained attribute representations followed by a score layer to make the prediction. The three tasks and their uniquely associated layers are described in detail in the below part.

3.1.1.2 Main Triplet Ranking Task

Our main task is sketch-photo ranking, and in this respect our network is similar to the state-of-the-art triplet network used in [3], except for the additional dropout to reduce overfitting. The main task is trained by supervision in the form of triplet tuples, with each instance tuple $\{s, p^+, p^-\}$ containing an anchor sketch s, positive photo p^+ and negative photo p^-. Corresponding to these input elements, the network has three branches and the goal is to learn a representation, such that the positive photo p^+ is ranked above the negative photo p^- in terms of its similarity to the query sketch s. To this end, the main task loss function is triplet ranking loss:

$$L_\theta\left(s, p^+, p^-\right) = \max\left(0, \Delta + D\left(f_\theta\left(s\right), f_\theta\left(p^+\right)\right) - D\left(f_\theta\left(s\right), f_\theta\left(p^-\right)\right)\right) \tag{3.1}$$

where θ represents the parameters of the network, $f_\theta(\cdot)$ denotes the learned deep feature of the corresponding network branch, $D(\cdot, \cdot)$ denotes the squared Euclidean distance, and Δ is the required margin in the triplet loss.

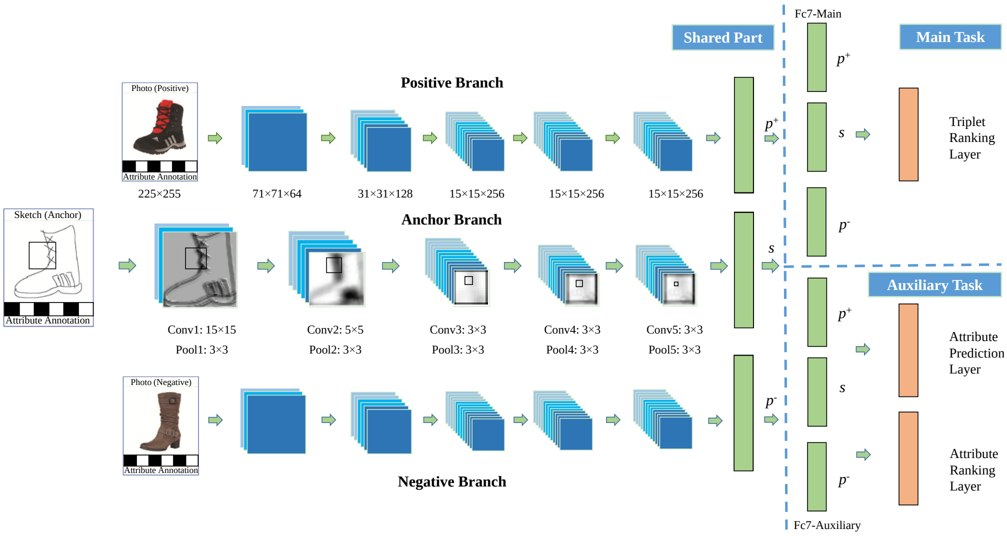


Figure 3.1: Network architecture of the proposed deep multi-task fine-grained SBIR model.

3.1.1.3 Attribute Prediction Task

In order to encourage the learned network representation to encode semantically salient properties of objects (and thus help the main task to make better (dis)similarity judgements for ranking), we let the network be aware of the semantic information by requiring the network to predict semantic attributes – such as whether a shoe is with shoelace, or whether a chair has wheels. For this task we assume that each training sketch s (or photo p) is annotated with N different semantic attributes, thus providing training tuples $\{s, t_1^s \dots t_N^s\}$, where t_i^s denotes the ith attribute for the sketch s, with $1 \leq i \leq N$. Prediction of the attribute vector of the input sketch and photo is thus a multi-label classification problem because attributes are not mutually exclusive. For convenience, we assume that each attribute is binary. Although some attributes can be correlated, this is not a limitation of our framework as our attribute prediction branch can still deal with most general case. The attribute prediction loss in our framework is the cross-entropy between the attribute labels and predictions $f_\theta^{ap}(\cdot)$, so for sketch attribute prediction we can have

$$L_p\left(s,t^s\right)=-\frac{1}{N}\sum_{n=1}^{N}\left[t_n^s\log f_{\theta,n}^{ap}\left(s\right)+\left(1-t_n^s\right)\log\left(1-f_{\theta,n}^{ap}\left(s\right)\right)\right], \tag{3.2}$$

and similarly the loss functions for the positive and negative photos are obtained by replacing s with p^+ and p^- respectively. This attribute prediction task can then be trained simultaneously with the main sketch-photo ranking task. Note that the attribute prediction task is considered in both photo and sketch branch, indirectly helping the two domains align.

3.1.1.4 Attribute Ranking Task

The attribute-prediction task above ensures that the learned representation encodes semantically salient features that support attribute prediction. Since retrieval ranking is the main task, the attribute prediction would not be used during test-time. This task's effect on the main task is thus implicit rather than direct. However, as a semantic representation, attributes are domain invariant and thus intrinsically useful for matching a photo with a query sketch. To this end, we introduce a third task of attribute-level sketch-photo matching which matches sketch to photo based on the predicted attributes of sketch and photo input rather than on an internally generated representation.

The loss function used for this task deserves some thought. A straightforward choice would be treating the attribute prediction exactly the same way as the learned deep representations from the bottom five feature extraction layers of the network and use a loss that is similar to that in Eq. (3.1), *i.e.*, a triplet ranking loss. Specifically, since the attribute predictions are probabilities, we compare attribute predictions from the three branches with cross-entropy rather than squared Euclidean distance as in the main task:

$$L_a\left(s,p^+,p^-\right)=\max\left(0,\Delta+H\left(f_\theta^{ap}\left(s\right),f_\theta^{ap}\left(p^+\right)\right)-H\left(f_\theta^{ap}\left(s\right),f_\theta^{ap}\left(p^-\right)\right)\right), \tag{3.3}$$

where $H(\cdot)$ is the cross-entropy between the attribute prediction vectors of the corresponding branches. However, there is a subtle but critical difference between the learned deep feature representation and attribute predictions: they have very different dimensionalities – the attributes are in the order of 10s whilst the deep features are 1000s. This means that they have different levels of discriminative power and thus need to be treated differently when designing cross-domain matching losses. In particular, given a dozen of attributes, many similar photo images could have very similar or even identical sets of attributes; forcing them to be different in order to enforce the ranking as in Eq. (3.3) would be too strong a constraint that is difficult to meet. Taking this into consideration, a more relaxed attribute-similarity loss function is adopted instead:

$$L_a\left(s, p^+, p^-\right) = H\left(f_\theta^{ap}\left(s\right), f_\theta^{ap}\left(p^+\right)\right), \tag{3.4}$$

which forms a less strong constraint that the positive photo should have similar attributes to the anchor sketch, and is found to be empirically better than the full triplet version of attribute ranking loss in our experiments. This attribute similarity loss obviously has an effect on how the training tuples are selected, *i.e.*, the sampling strategy which will be discussed in Sec. 3.1.3.

3.1.1.5 Multi-Task Training

With the three tasks, the overall loss function for multi-task training of our network is given by a weighted sum in Eq. (3.5).

$$\begin{aligned} L\left(s, p^+, p^-\right) = {} & L_\theta\left(s, p^+, p^-\right) + \lambda_a L_a\left(s, p^+, p^-\right) + \lambda_s L_p\left(s, t^s\right) + \lambda_{p^+} L_p\left(p^+, t^{p^+}\right) \\ & + \lambda_{p^-} L_p\left(p^-, t^{p^-}\right) + \lambda_\theta \|\theta\|_2^2 \end{aligned} \tag{3.5}$$

where the first term is the main ranking task, the second term is the attribute ranking task, the next three are attribute predictions for anchor sketch branch, positive photo branch, and negative photo branch, respectively, and the last one is a regularization term to suppress the complexity of weights [75]. Here the relative weight of each side task is denoted by the hyper parameters $\lambda = \left(\lambda_a, \lambda_s, \lambda_{p^+}, \lambda_{p^-}\right)$.

3.1.1.6 Multi-Task Testing

At run-time the main and attribute-ranking tasks are used together to generate an overall similarity score for a given sketch/photo pair. All sketch/photo pairs are ranked, and the retrieval for a given sketch is the similarity-sorted list of photos. Specifically, for a given query sketch s the similarity to each image p in the gallery set is calculated as

$$R_s\left(s,p\right) = D\left(f_\theta\left(s\right), f_\theta\left(p\right)\right) + \lambda_a H\left(f_\theta^{ap}\left(s\right), f_\theta^{ap}\left(p\right)\right). \tag{3.6}$$

where $D(\cdot)$ and $H(\cdot)$ are squared Euclidean distance and cross-entropy respectively.

3.1.2 Staged Model Pre-training

A staged pre-training strategy is adopted similar to that of [3]. Specifically, first, a single branch classification model with the same feature extraction layers as the proposed full model is pre-trained to first classify ImageNet-1K data (encoded as edge maps). This model is very similar to the Sketch-a-Net model [31] designed for sketch classification. This is followed by fine-tuning on the 250 classes TU-Berlin sketch recognition task. After that, this single branch network is extended to form a three-branch Siamese triplet ranking network. Each branch is initialised as the pre-trained single-branch model, and the model is then fine-tuned on a category-level photo-sketch dataset re-purposed for fine-grained SBIR as in [3]. After these three stages of pre-training, the full model with two added auxiliary-tasks and the overall loss in Eq. (3.5) is then initialised and fine-

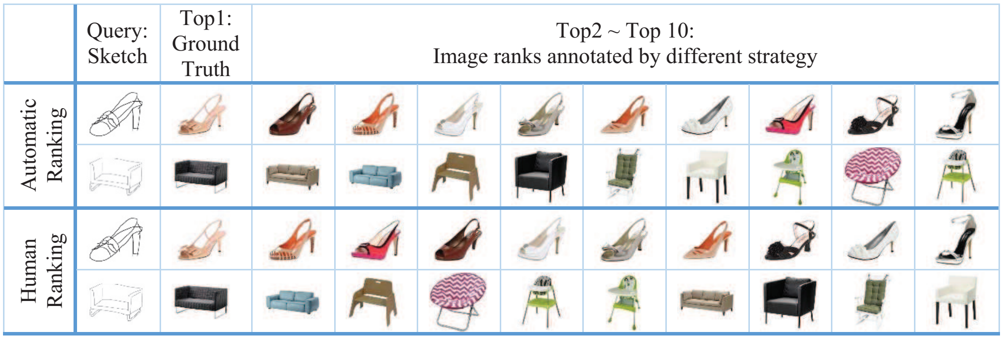


Figure 3.2: Rank lists generated automatically and by global ranking of human triplets.

tuned with the fine-grained SBIR dataset for instance-level sketch-based photo retrieval.

3.1.3 Attribute-based Sampling Strategy

Determining an optimal sampling strategy for constructing the anchor-positive-negative triplet tuples for model training is critical. There two major choices: (1) how to generate the triplets and (2) how to select a subset of them for model training. For the former, one straightforward choice is that given each anchor/query sketch, to form exhaustive photo pairs and present the resultant triplets for humans to annotate which photo is more similar to the anchor. However, this is intractable even for a moderate data size. Hence in [3] the top-10 ranked photos for a given anchor is selected, where exhaustive human annotation is collected, yielding a total of $10 \cdot 9/2 = 45$ triplets per sketch. All such superset of 45 human annotated triplets are then used to train a triplet ranking model. However, there are two problems: (1) even with pre-screening, the exhaustive annotation is still expensive, and (2) the collected annotations are error-prone, since top ranked photos are all very similar to each other, making triplet ranking a challenging task for humans to perform reliably (see Figure 3.2 – some pairs in the list are hard to order by similarity with respect to the query). The reliability of human annotation can be improved by employing a global ranking method such as [76] to correct annotation noise. However, there is no solution to the scalability issue. In this work, a new way to generate the triplets and a novel sampling strategy are developed, which entirely removes the need for the otherwise non-scalable and unreliable human triplet annotations.

3.1.3.1 Triplet Generation

Instead of choosing top-10 most similar photos and asking humans to annotate (as in [3]), we automatically generate triplets based on a strict top-10 ranking induced by attribute and feature similarity. More specially, we first use attribute similarity to construct a top-10 candidate list of most similar photos given a query sketch. ImageNet CNN features are then used to further rank these photos by similarity with respect to the ground-truth match. Intuitively this strategy can be seen as using semantic attribute properties to generate a meaningful short list, but otherwise driving the cross-domain ranking objective by more subtle photo-photo similarity encoded by a well-trained ImageNet CNN. It follows that a total of 45 triplets can be automatically generated by enforcing ranks among candidate photos within each triplet (*i.e.*, photo with higher rank is annotated as positive and vise versa). In Figure 3.2, we compare our automatic top-10 ranking with a globally optimised ranking computed from human triplets [76]. Overall the automatic one is of comparable (or better) quality than the more costly manually generated list. Another motivation behind this well-trained ImageNet CNN based triplet generation is to transfer the knowledge of ranking in the same domain from a well-trained heavier model to our model. Since well-trained ImageNet CNN are super informative to differentiate thousands of classes and the ranking is generated on the same domain avoiding ranking them with gap cross sketches and photos, we believe the knowledge of this generated ranking is more reliable and suitable to be transferred.

3.1.3.2 Triplet Sampling

The second novel feature is that instead of using all 45 triplets as per [3], we sample the 9 hardest ones for model training, each consisting of the anchor and two photos of *neighbouring ranks* (*e.g.*, anchor-R1-R2 or anchor-R4-R5). We show empirically that this choice of learning curriculum significantly boosts model performance compared to alternatives ranging from exhaustive sampling, easy, and medium. Seemingly counter-intuitive to the conventional 'more data is better' maxim, there are two explanations of why sampling a small subset of hard samples helps: (a) After extensive (three) stages of model pre-training, the model has already learned a strong domain-invariant representation; it is therefore 'ready' to accept hard training samples [77]. (b) Importantly, the introduction of the two additional attribute-based side tasks means that the model is much more robust against overfitting with small training data size. (c) In addition, the model is seeing the same data in both cases, but in our sampling strategy, the model is able to focus more on some hard comparisons.

3.2 Experiments

3.2.1 Datasets and Settings

3.2.1.1 Training and Evaluation Data

We use the same shoe and chair FG-SBIR datasets introduced by [3]. For training, 304 sketch-photo pairs of shoes, and 200 pairs of chairs are used. Each sketch/photo comes with attribute annotations, which are used to obtain the top 10 photo rank list in [3] and additionally to learn attribute-based tasks in our multi-task model. Data augmentation like flipping and cropping is applied.

3.2.1.2 Network Implementation

We use the Caffe library [78] to implement our deep multi-task model. Task-importance parameters are set to $\lambda = \left(\lambda_a, \lambda_s, \lambda_{p^+}, \lambda_{p^-}\right) = \{1, 0.01, 0.01, 0.01\}$, *i.e.*, the main and attribute-level ranking tasks have equivalent weight, and the attribute-prediction tasks all have the same lower weights, considering that the classification task is a much easier task compared to the ranking task. The single loss margin is set to $\Delta = 1$. During joint training, the batch size is set to 128, and the network is trained with a maximum of 25000 iterations. The base learning rate is 0.001 and weight decay (λ_θ) is set to 0.0005.

3.2.1.3 Evaluation metrics

To evaluate performance, we use the same two evaluation metrics as [3, 79]: Top K retrieval accuracy for $K = 1$ and $K = 10$. This corresponds to the use scenario where there is a particular object that the user needs to retrieve exactly. An alternative scenario, is where the user just wants to see similar items to the sketch, and in this case the overall ordering is the salient metric. For this we use % of correctly ranked triplets, which reflects how well the predicted triplet ranking agrees with that of humans.

3.2.1.4 Baselines

We compare our multi-task model with several baselines, including the state-of-the-art fine-grained instance-level triplet ranking [3] (termed as the Triplet model), which is also the main deep alternative to the proposed framework. As representatives of the classic approaches, RankSVM is trained base on HOG features extracted and encoded as either bag of words (termed as BoW-HOG+rankSVM), or large dense vectors (termed as Dense-HOG+rankSVM). As representatives of alternative deep feature-based approaches, we also extract Sketch-A-Net deep features [31], and 3D shape deep features [33] for RankSVM training (termed as 3DS Deep+RankSVM and ISN Deep+RankSVM respectively).

3.2.2 Results

3.2.2.1 Comparisons against the State-of-the-art

FG-SBIR retrieval performance is first evaluated to compare our multi-task model with the state-of-the-art methods outlined previously. From the results in Table 3-A we can see that our MTL obtains much higher accuracy compared to previous work, especially for Rank 1 (Top 1) matching accuracy – around 10% improvements over the state-of-the-art in [3] are achieved, despite the fact that the triplet model in [3] requires costly human triplet annotations which are not used by our framework.

Table 3-A: Comparative results against state of the art retrieval performance.

Shoe Dataset	Top 1	Top 10	Trip Acc	Chair Dataset	Top 1	Top 10	Trip Acc
BoW-HOG + rankSVM	17.39%	67.83%	62.82%	BoW-HOG + rankSVM	28.87%	67.01%	61.56%
Dense-HOG + rankSVM	24.35%	65.22%	67.21%	Dense-HOG + rankSVM	52.57%	93.81%	68.96%
ISN Deep + rankSVM	20.00%	62.61%	62.55%	ISN Deep + rankSVM	47.42%	82.47%	66.62%
3DS Deep + rankSVM	5.22%	21.74%	55.59%	3DS Deep + rankSVM	6.19%	26.80%	51.94%
Triplet model [3]	39.13%	87.83%	69.49%	Triplet model [3]	69.07%	97.94%	72.30%
Ours	**50.43%**	**91.30%**	**70.59%**	Ours	**78.35%**	**98.97%**	**73.13%**

3.2.2.2 Contributions of Auxiliary Tasks

The first ablation study investigates the contributions of different auxiliary tasks. The main reason our MTL model outperforms the-state-of-the-art is due to the benefit provided by the auxiliary attribute-related auxiliary tasks: indirectly in the case of attribute prediction (AP) and directly in the case of attribute ranking (AR). To demonstrate this we compare the performance of our full model with the performance obtained by removing one or both of the auxiliary tasks (*e.g.*, "Ours - AP" means our full model with the AP task removed). From the results in Table 3-B, we can see that each task helps, as the performance drops when either auxiliary task is removed, and drops further when both of them are removed. Notice that the hyper-parameters for the importance of auxiliary tasks is set by experience value. A too higher or lower may lead to different tasks converge at a different time.

Table 3-B: Contribution of the proposed attribute side tasks.

Shoe Dataset	Top 1	Top 10	Trip Acc	Chair Dataset	Top 1	Top 10	Trip Acc
Ours - AP - AR	37.39%	82.61%	66.57%	Ours - AP - AR	50.52%	91.75%	69.62%
Ours - AR	45.22%	87.83%	**72.37**%	Ours - AR	72.16%	98.97%	72.00%
Ours - AP	44.35%	86.96%	71.34%	Ours - AP	72.16%	98.97%	72.10%
Ours	**50.43**%	**91.30**%	70.59%	Ours	**78.35**%	**98.97**%	**73.13**%

3.2.2.3 Comparison of Triplet Generation and Sampling Strategies

We investigate two ways of generating triplets and various sampling strategies in this section. Generation: the triplets are generated either automatically (using attribute/feature ranking) or manually by humans. As mentioned earlier, the original human annotation can be noisy, we therefore clean human annotations by inferring a globally optimised rank list from the annotated pairs using the generalised Bradley-Terry model [76]. Sampling: using either generation method, 10 photos are ranked for any given sketch which gives a total of $10 \cdot 9/2 = 45$ triplets. Sampling options include: (i) *Exhaustive*: use all 45 triplets with no sampling, or (ii) *Hard*: sample the 9 hardest triplets as proposed. We also train a network using the same human annotated triplets used by [3] as baseline.

Table 3-C: Impact of different triplet annotation strategies.

Method	Shoe Dataset			Chair Dataset		
	Top 1	Top 10	Trip Acc	Top 1	Top 10	Trip Acc
Auto-generated (exhaustive)	43.48%	86.09%	70.38%	68.04%	97.94%	70.58%
Auto-generated (hard only)	**50.43**%	**91.30**%	70.59%	**78.35**%	98.97%	73.13%
Human-optimised (exhaustive)	43.48%	87.83%	70.88%	71.13%	98.97%	73.29%
Human-optimised (hard only)	47.83%	87.83%	70.28%	77.32%	**100.00**%	**73.95**%
Human original (as in [3])	42.61%	89.57%	**71.29**%	71.13%	100.00%	73.84%

Table 3-C compares results obtained by our model using different triplet generation/sampling strategies. We can draw the following conclusions: (1) Our automatically generated hard triplet sampling strategy performs the best overall. (2) In general, using a smaller number of 9 hard triplets performs better than the 45 exhaustive triplets, for either manual or automatic generation. This suggests that hard triplets help learn a better fine-grained cross-domain representation. (3) Overall, the auto-generated triplets produce better performance than the human annotated triplets. The above results are somewhat surprising, as the conventional wisdom is that 'more data is always better'

Table 3-D: The influence of training triplet difficulty on testing performance.

Shoe Dataset	Top 1	Top 10	Trip Acc	Chair Dataset	Top 1	Top 10	Trip Acc
Easy triplets	39.13%	80.87%	70.24%	Easy triplets	69.07%	96.91%	68.75%
Medium triplets	41.74%	86.09%	**71.05%**	Medium triplets	68.04%	97.94%	71.75%
Hard triplets	**50.43%**	**91.30%**	70.59%	Hard triplets	**78.35%**	**98.97%**	**73.13%**

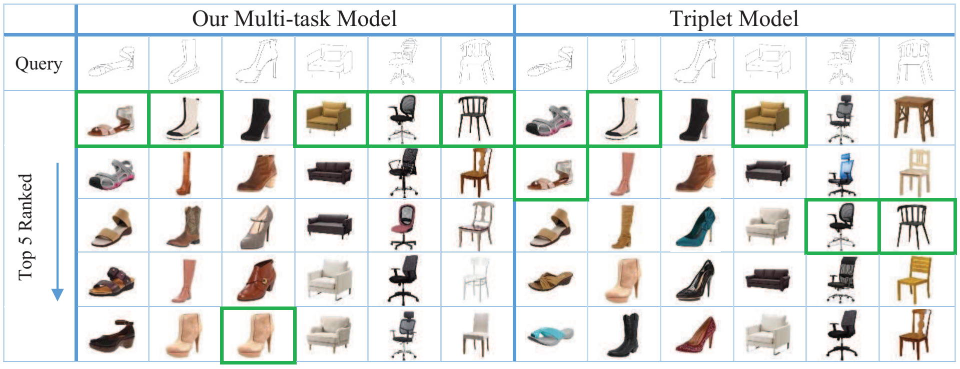


Figure 3.3: Retrieval results of our proposed method, compared with that of [3].

and that careful manual annotation should be better than automatic annotation. We attribute the superiority of fewer harder triplets to the fact that the base model is already quite well pre-trained, so that at the point we start training it is 'ready' for difficult examples, in a curriculum learning sense [77]; and the superiority of generated triplets to manually annotated triplets to the fact that the similarity judgements are quite hard to make reliably given the short list of similar images, so in this case the human annotation is no more reliable than the automatic annotation.

We next investigate further the issue of sampling triplets according their difficulty level. We define hard triplets as before, where each triplet spans a distance of 1 on the rank list. Medium triplets are defined as those with distance 2 and 3, and easy triplets are those with distance larger than 3. Thus within the top 10 list, the 45 exhaustive triplets include 9 hard, 15 medium and 21 easy ones. The results in Table 3-D show that performance increases with triplet difficulty, supporting our hypothesis that hard triplets are the most valuable at this stage.

3.2.2.4 Qualitative Results

Example retrieval results of our proposed multi-task model are shown in Figure 3.3, where the retrieved image with green box is the ground truth. From the qualitative results, we can see that our method can achieve satisfying result, and can find out the ground-truth match in the rank 1 place in most cases, while the baseline model can only guarantee less than half of the cases. Moreover, in the high-heel shoe test case, our model can also correctly retrieve the true-match shoe, and returned it in the top 5 result, but the baseline model fail to obtain the correct photo.

3.2.2.5 Computational Cost

Our deep multi-task model is trained on an Nvidia Tesla K80 GPU. The re-implementation of the sketch triplet model takes about 5 days, as detailed in [3]. The joint training of the proposed deep multi-task model takes about 7 hours for 25,000 iterations of batches for either chair or shoe dataset. In the testing stage, the time cost on retrieval in the shoe's and chair's gallery per query is about 0.22s and 0.18s, respectively.

3.3 Summary

In this chapter, we introduce a deep multi-task attribute-based model for fine-grained SBIR. By constructing attribute-prediction and attribute-based ranking side-tasks alongside the main sketch-based image retrieval task, the main task representation is enhanced due to being required to encode semantic attributes of sketches and photos, and moreover the attribute predictions can be exploited to help make similarity predictions at test time. The combined result is that performance is significantly improved compared to previous the state-of-the-art using a deep triplet ranking task alone. Beyond this we showed that somewhat surprisingly the human subjective triplet annotation is not be critical for obtaining good performance. This means that it is relatively easy to extend the method to new categories and larger datasets, since attribute annotation grows only linearly rather than cubically in the amount of data.

Chapter 4

Fine-grained Sketch-based Image Retrieval with Text

Semantic information is also expressed in sketches and can be exploited with the fine-grained information in the sketch recognition and retrieval tasks. Specifically, fine-grained image retrieval (FGIR) [14, 31, 80] 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. It is arguably a more difficult problem when compared with fine-grained categorisation [81, 82] for that (i) it seeks intra-category ranking other than basic categorisation, and (ii) retrieval is often conducted cross-modal, *e.g.*, sketch/text as input modality, as oppose to within the single photo modality. 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 photo retrieval. For example, using the keyword 'shoe' in a Google/Bing image search engine generates very satisfying results - the first few return pages all contain shoe images. However, when it comes to instance-level 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 [3, 14]. Human sketches have been advocated by many as a natural input modality since it implicitly captures both fine-grained appearance and holistic structure information [3, 14]. A sketch is perhaps worth one hundred words but takes much less efforts to produce. With the popularity of touch-screen devices, drawing sketches 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 4.1 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.

Prior work on fine-grained image retrieval mainly investigate using sketches as input. They primarily focus on closing the semantic gap between the two modalities while completely ignoring the text modality. Specifically, state-of-the-art fine-grained sketch-

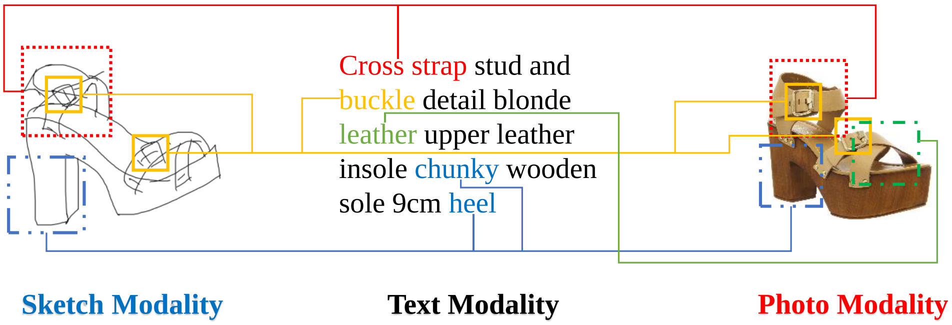


Figure 4.1: Relationship between sketch, photo, and text modalities.

based image retrieval models [3, 14] adopt a multi-branch deep convolutional neural network (CNNs). Each modality has a corresponding branch which consists of multiple convolutional/pooling layers followed by pairwise verification or triplet ranking losses to align the domains on the FC layers. There also exists plenty of work on aligning text and photo for cross-modal retrieval. Again, all of them employ a two-modality setting and use probabilistic model [83], metric learning [84], or subspace learning [85] to link the two modalities. We differ from all previous text-photo cross-modal work on two key aspects: (i) we employ full text descriptions similar to those studied in image captioning [86], which is more fine-grained than tags/keywords found in popular text-photo cross-modal datasets, and (ii) we conduct intra-category instance-level retrieval whereas all previous work was designed and evaluated on category basis. Over and above all, none of the previous literature addressed learning a unified embedded space for all three modalities (text, photo and sketch), and investigated if text and sketch can complement each other in the fine-grained retrieval setting.

In this chapter, 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 [3, 14]. 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.

4.1 Methodology

In this section, we proposed a multi-modal deep network to achieve fine-grained retrieval among sketch, text and photo modalities. We first introduce the whole network architecture, and then go through the multi-modal feature embeddings and losses used to align them.

4.1.1 Network Architecture

The architecture of our model is shown in Figure 4.2. 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 [3]. Specifically, the backbone part of sketch branch and photo branches share the network parameters, and are configured with 5 consecutive convolutional layers with

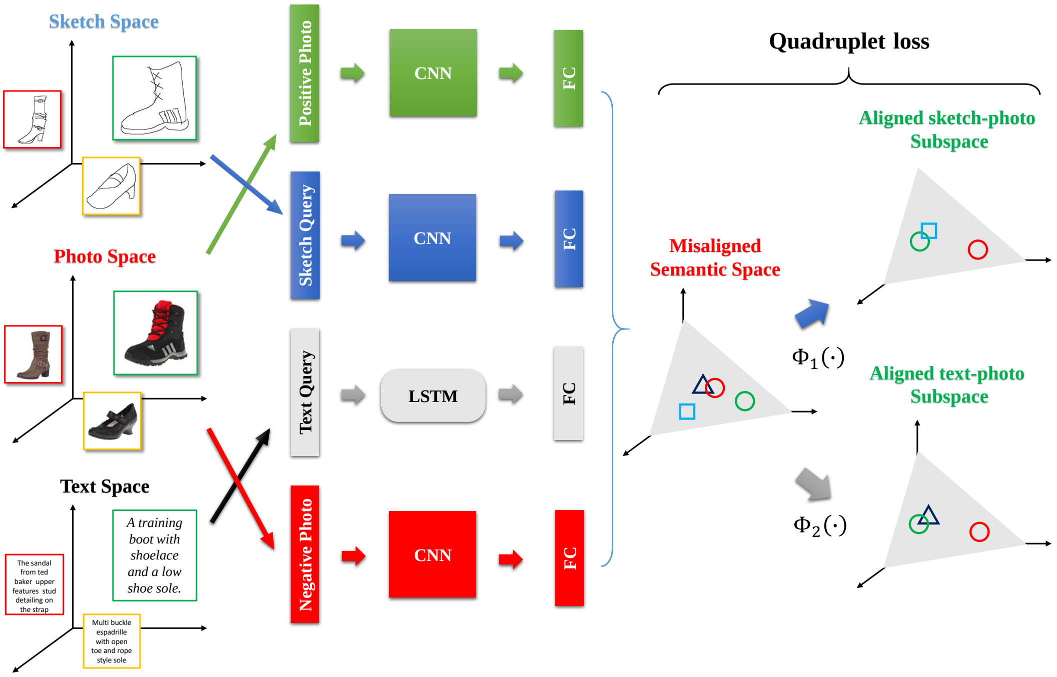

Sketch Space
Photo Space
Text Space
A training boot with shoelace and a low shoe sole.
The sandal from ted baker upper features stud detailing on the strap
Multi buckle espadrille with open toe and rope style sole
Positive Photo
Sketch Query
Text Query
Negative Photo
CNN
CNN
LSTM
CNN
FC
FC
FC
FC
Quadruplet loss
Misaligned Semantic Space
Aligned sketch-photo Subspace
Aligned text-photo Subspace
$\Phi_1(\cdot)$
$\Phi_2(\cdot)$

Figure 4.2: Architecture of the proposed multi-modal learning framework.

ReLU activation and 2 fully-connected layers afterwards also with ReLU activation, following the state-of-the-art FG-SBIR approach [3].

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. Experimental results show that the proposed framework has outperform all the baselines, including the state-of-the-art alternative [3].

4.1.2 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 [3] are constructed, where we use Siamese convolution and pooling layers with weights tied among different domains. Then fully-connected layers are applied to reduce the high dimensional convolution layer feature to a lower dimensional feature space, while the following fully-connected layer weights project the embedding from each modality to the shared latent space.

We also apply the shared/pairwise dropout strategy [87] 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.

4.1.3 Fine-grained Text Description Embedding

It is natural to utilise 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). The encoded vectors from different timestamps are then sent to the bidirectional LSTM network to train the cells. Our dynamic LSTM cells are then updated following [88]. 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 [89] as the overall representation of the input sentence, *i.e.*, the text embedding.

4.1.4 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^-, represent 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,

$$\begin{aligned} L\left(s, t, p^+, p^-\right) = &\max\left(0, \Delta + D\left(\Phi_1(f(s)), \Phi_1(f(p^+))\right) - D\left(\Phi_1(f(s)), \Phi_1(f(p^-))\right)\right) \\ &+ \max\left(0, \Delta + D\left(g\left(t\right), \Phi_2(f(p^+))\right) - D\left(g(t), \Phi_2(f(p^-))\right)\right) \end{aligned} \tag{4.1}$$

where $g(t)$, $f(s)$, $f(p^+)$ and $f(p^-)$ denote the learned anchor text, anchor sketch, the positive and negative photo embedding, respectively. $D(\cdot, \cdot)$ is the distance metric, here we take the l_2 normalised squared euclidean distance 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 eliminate the domain gap among sketch, photo and text modalities. For example,

$$\Phi_1(f(p))) = W_1^\top f(p) + b_1 \tag{4.2}$$

, 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 photo-sketch, or text-sketch ranking/retrieval using the corresponding distance from the objective:

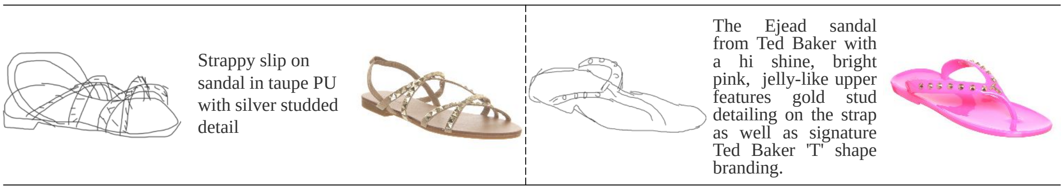

Strappy slip on sandal in taupe PU with silver studded detail
The Ejead sandal from Ted Baker with a hi shine, bright pink, jelly-like upper features gold stud detailing on the strap as well as signature Ted Baker 'T' shape branding.

Figure 4.3: Example of the shoe multi-modal dataset.

$$R_s\left(s,p\right) = -D\left(\Phi_1(f(s)),\Phi_1(f(p))\right) \tag{4.3}$$

$$R_s\left(t,p\right) = -D\left(g\left(t\right),\Phi_2(f(p))\right) \tag{4.4}$$

4.2 Fine-grained Multi-modal Retrieval Dataset

We contribute a new dataset for multi-modal learning, especially for fine-grained cross-modal retrieval. We collected 1374 sketch-photo-text triplets for shoes. Specifically, we collect shoe photos (in side view) and their corresponding descriptions from an online shopping website. After that, we ask volunteers to draw free-hand sketch for each given photo. Some examples are given in Figure 4.3. The collected fine-grained multi-modal dataset is available from `http://www.eecs.qmul.ac.uk/~js327`.

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 1,112 sketches, text descriptions and photos to train our multi-modal deep neural network, and 262 instances for testing. The dataset sounds small, but they can form much more quadruplets, and with proper data augmentation strategy, the data size problem is much more relieved.

4.3 Experiments

4.3.1 Implementation Details

4.3.1.1 Data Preprocessing and Augmentation

We pre-process the photos into edge maps via EdgeBox [90]. We then do random crop and random flip on both sketches and photos to enrich the training data, which is also a common data augmentation strategy. Similar to other preprocessing strategies in text modality, we remove all the stop words and symbols in the raw text description, as well as some rare words with maximum word count less than 5.

4.3.1.2 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 [3]. More specially, we first pre-train the sketch and photo branches on TU-Berlin dataset [1] and extracted edges from ImageNet [3], respectively. For the text branch in our model, we first use gensim word2vec model [91] (pretrained on Google News dataset) on our training text description to pre-train the word embedding on our network. After the pre-training stage, we then fine-tune our model on the collected dataset, with batch size 128. We choose stochastic gradient descent (SGD) as the optimizer to train our multi-modal model, with a learning rate at 0.0001. Moreover, to prevent over-fitting, we add the dropout layer on the fully-connect layer and the bidirectional LSTM cells at a 0.5 dropout rate, and we also put an l_2 regulariser with 0.0005 weight decay to reduce over-fitting.

4.3.1.3 Sampling Strategy

Our quadruplet sampler is inspired by sampling strategy in [92]. We stick the ground-truth photo as the positive instance to the anchor sketch and the anchor text, and select

100 nearest neighbour photos in the VGG feature space [93] as the hard negative samples for each sketch and corresponding text and ground-truth photo. Therefore, before data augmentation, 111,200 quadruplets are generated for the training stage. The benefit of our sampling strategy is that expensive human annotation on exhaustive generated pairs can be saved though this may or may not give a more accurate ranking considering that it is hard to distinguish between very similar instances.

4.3.2 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 further analysis to show insight about modality alignment, and how sketch-photo and text-photo can benefit each other.

4.3.2.1 Comparative Results against Baselines

Three baselines are chosen for comparison. The first is multi-view shallow CCA [94]. To obtain the multi-modal representations for this multi-view CCA framework, Sketch-a-Net 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 [95], one hidden fully-connected layer with 256 dimension (256D) transformed the same deep representation, and the CCA layer (32D) followed then 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 losses to match the embedding between sketch and photo, and between text and photo. The results in Tab. 4-A demonstrate that our proposed method are clearly superior to the other baselines. It is also interesting to find that both shallow CCA loss and deep CCA are not suitable for fine-grained retrieval, compared to the L2 loss and our unified quadruplet loss, as there are many images in our gallery are with high similarity, *i.e.*, already highly correlated.

Table 4-A: 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[94]	0.38%	4.20%	0.76%	4.58%
Deep CCA[95]	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%**

4.3.2.2 Benefit from Each Cross-Modal Learning

Our multi-modal learning model can also be viewed as a multi-task learning model, which has been proven useful in many computer vision problem. 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. Moreover, the side tasks can also benefit each other, as different side tasks in return provide more data to the shared latent tasks. Thus a better shared representation is learned from more data and help to improve the performance of both tasks.

In our multi-modal framework, one task is the fine-grained sketch-to-photo retrieval, while the other one 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 Tab. 4-B.

Table 4-B: 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%**

4.3.2.3 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. This is also to show the effectiveness of the proposed sketch-to-photo retrieval module. Here we compared our sketch-photo model with two most recent state-of-the-arts: triplet Sketch-a-Net model [3] and triplet GoogleNet [14]. The results in Tab. 4-C 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 4-C: Performance comparison on fine-grained SBIR.

Model	Top 1	Top 10
Triplet Sketch-a-Net[3]	46.56%	82.82%
Triplet GoogleNet[14]	45.42%	79.77%
Our sketch-photo model	49.24%	82.06%
Our full model	**50.38%**	**84.73%**

4.3.2.4 Photo-text Embedding Alignment Performance

We evaluate against the captioning approach to FG-TBIR by applying the CNN-RNN architecture as detailed in [86] to generate descriptions for our gallery photos, 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 Tab. 4-D, 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-D: Photo-text embedding alignment performance with different methods.

Model	Top 1	Top 10
Shallow CCA[94]	0.38%	5.34%
Deep CCA[95]	3.05%	18.70%
Photo caption model[86]	7.60%	24.40%
Our text-photo model	8.78%	33.97%
Our full model	**12.60%**	**37.40%**

4.3.2.5 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. 4.4, where the ground-truth photo is highlighted using a green bounding box. Note that even though all the photo are left to right oriented, our model is robust to the view point change, as we do data augmentation by random horizontal flip.

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. 4.5.

4.3.2.6 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. As a result, 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. 4.6 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

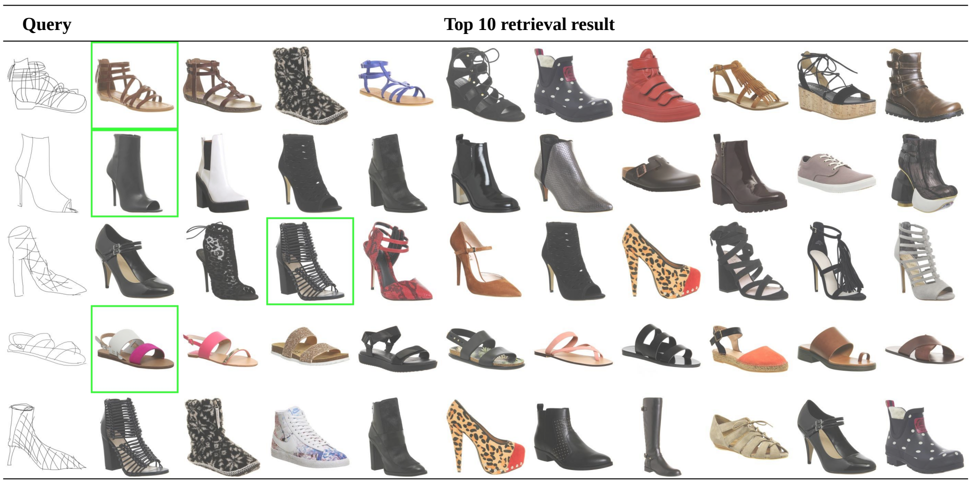
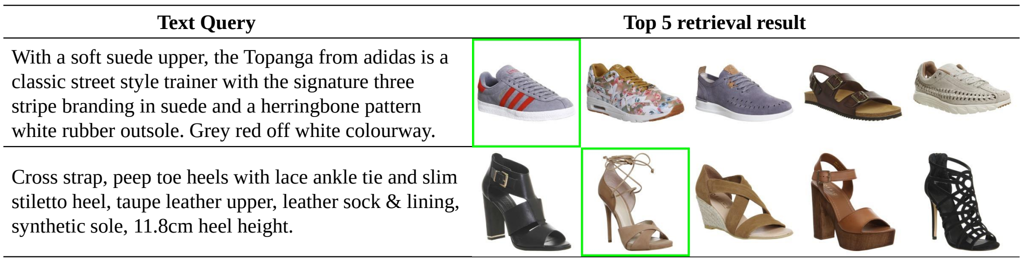


Figure 4.4: Example of fine-grained sketch-based image retrieval.



With a soft suede upper, the Topanga from adidas is a classic street style trainer with the signature three stripe branding in suede and a herringbone pattern white rubber outsole. Grey red off white colourway.

Cross strap, peep toe heels with lace ankle tie and slim stiletto heel, taupe leather upper, leather sock & lining, synthetic sole, 11.8cm heel height.

Figure 4.5: Example of fine-grained text-based image retrieval.

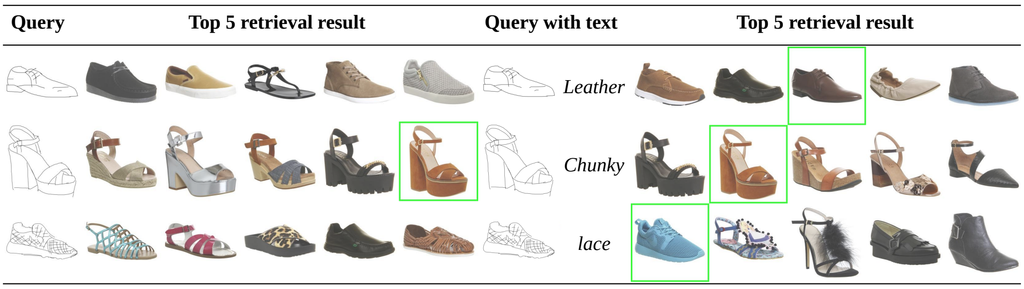


Figure 4.6: Example of fine-grained image retrieval with both sketch and text query.

not convey features like material and fine texture, which are however straightforward to describe in text. Table 4-E shows that after fusing the two query modalities, the retrieval performance is improved compared to that obtained using each modality alone. This also suggests that our model can exploit the complementarity of the two modalities for better retrieval performance.

Query	Model	Top 1 acc	Top 10 acc
sketch → photo	Our full model	50.38%	84.73%
text → photo	Our full model	12.60%	37.40%
(sketch + text) → photo	Our full model	**52.67**%	**87.02**%

Table 4-E: The performance of fine-grained image retrieval when both sketch and text is available as input.

4.4 Summary

In this chapter, 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.

Chapter 5

Photo-to-Sketch Synthesis

Temporal information is one specific feature for sketch modality, which is not possessed by the photo modality. The temporal information therefore distinguishes the sketching process with imaging process and required specific computer vision processing to handle. The sketching process is also the key to understand human visual processing. Think about this question: what do we see when our eyes perceive a grid of pixels from a real-world object? We can quickly answer this question by sketching a few line strokes. Despite the fact that drawings like this may not exactly match the object as captured by a photo, they do tell us how we perceive and represent the visual world around us, that is, we as humans convey our perception of objects abstractly but semantically.

In this context, it is natural to ask to what extent a machine can see. For decades, researchers in computer vision have dedicated themselves to answering this question, by injecting intelligence and supervision into the machine with the hope of seeing better. This is mostly done by formulating several specific constrained problems, such as classification, detection, identification, and segmentation.

In this chapter, we take one step forward – teaching a machine to generate a sketch

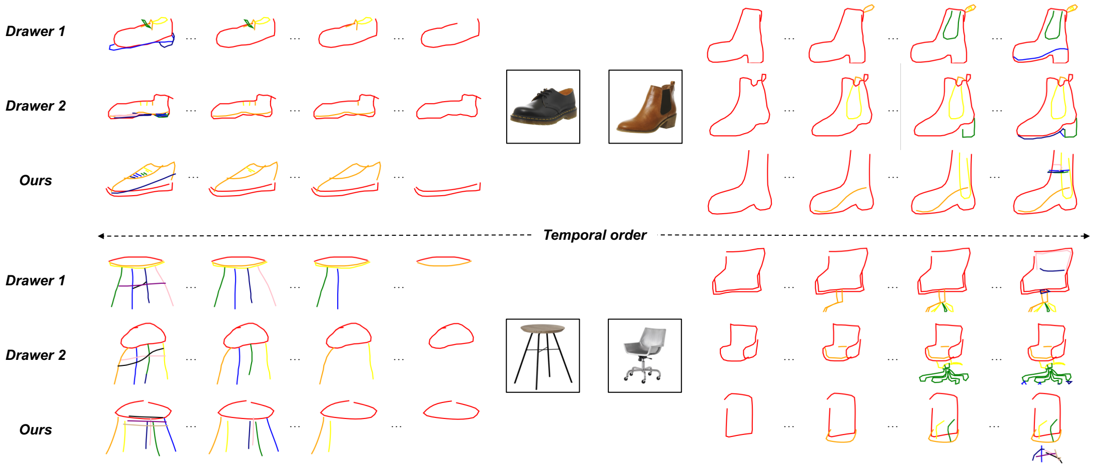


Figure 5.1: Given one object photo, our model learns to sketch stroke by stroke, abstractly but semantically, mimicking human visual interpretation of the object. Our synthesised sketches maintain a noticeable difference from human sketches rather than simple route learning (*e.g.*, shoelace for top left shoe, leg for bottom right chair). Photos presented here have never been seen by our model during training. Temporal strokes are rendered in different colours. Best viewed in colour.

from a photo just like humans do. This requires not only developing an abstract concept of a visual object instance, but also knowing what, where and when to sketch the next line stroke. Figure 5.1 shows that the developed photo-to-sketch synthesizer takes a photo as input and mimics the human sketching process by sequentially drawing one stroke at a time. The resulting synthesised sketches provide an abstract and semantically meaningful depiction of the given object, just like human sketches do.

Photo-to-sketch synthesis can be considered as a cross-domain image-to-image translation problem. Thanks to the seminal work of [2, 96], we are able to construct a generative sequence model with recurrent neural network (RNN) acting as a neural sketcher. However, the synthesised sketches are not conditional on specific object photos. To address this problem, one can encode the photo via a convolutional neural network (CNN) and feed the code into the neural sketcher. Such a photo-to-sketch synthesizer essentially follows the traditional encoder-decoder architecture (see Figure 5.2(a)), and has been taken by most existing image-to-image translation models [4, 97]. Training such

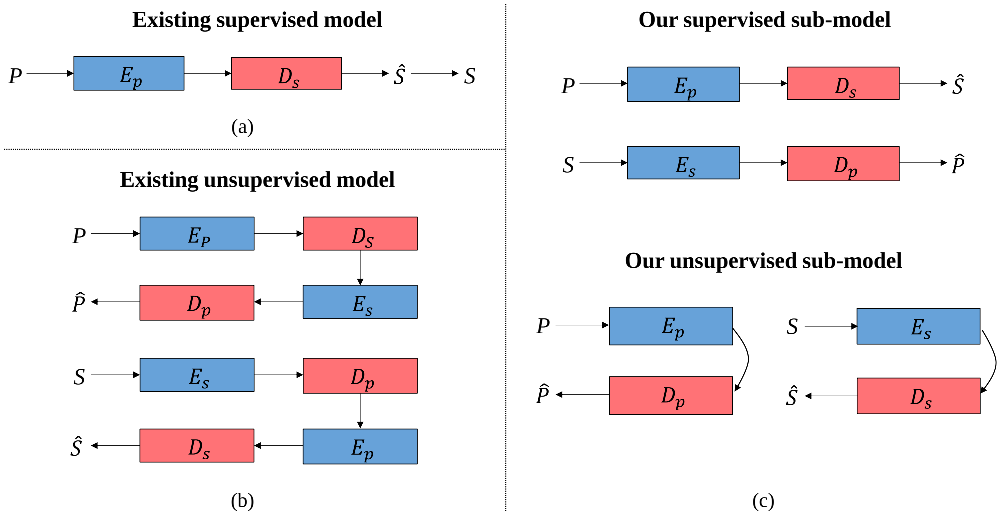


Figure 5.2: (a) Existing supervised image-to-image translation framework, where mapping is one-way only. (b) Existing unsupervised image-to-image translation models enforce cycle consistency to address the highly under-constrained one-to-one mapping problem. (c) Our supervised-unsupervised hybrid model with dual/two-way supervised translation sub-models and two unsupervised sub-models with shortcut cycle consistency. This takes advantage of the noisy supervision signal offered by photo-sketch pairs, as well as learning from within-domain reconstruction.

a model is done in a supervised manner requiring cross-domain image pairs: in our problem, these are photo-sketch pairs containing the same object instances. Compared to image-to-image translation, the key challenge for learning instance-level photo-to-sketch synthesis is that training pairs provide highly noisy supervision: Different sketches of the same photo have large style and abstraction differences between them (see Figure 5.3). This makes our problem highly noisy and under-constrained.

In order to achieve photo-to-sketch synthesis under noisy photo-sketch pairs as supervision, we address the limitations of existing cross-domain image translation models by proposing a novel framework based on multi-task supervised and unsupervised hybrid learning (see Figure 5.2(c)). Taking an encoder-decoder architecture, our primary task is $D(E(photo)) \to sketch)$ where a photo is first encoded by E and then decoded into a sketch by D. To help learn a better encoder and decoder, we introduce the inverse problem $(D(E(sketch)) \to photo)$ so that the supervised model learning can be done

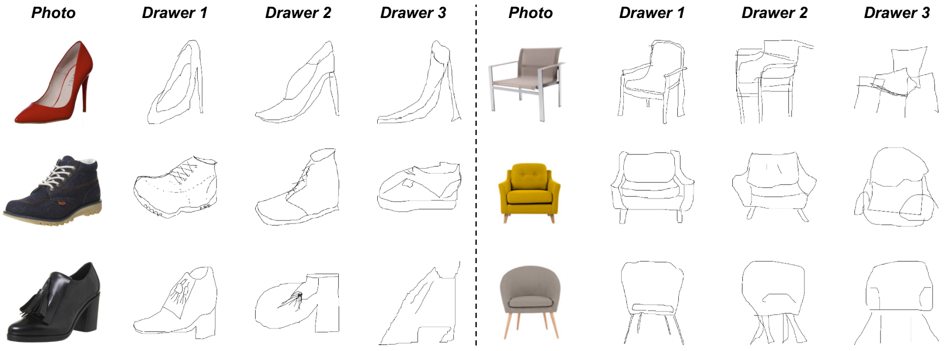


Figure 5.3: Given a reference photo, sketches drawn by different people exhibit large variation in style and abstraction levels. Some of them are poor in depicting the object instances in the corresponding photos.

in both directions. Importantly, we also introduce two unsupervised learning tasks for within-domain reconstruction, *i.e.*, $D(E(photo)) \rightarrow photo$ and $D(E(sketch)) \rightarrow sketch$. This hybrid learning framework differs significantly from existing approaches in that: (1) It combines supervised and unsupervised learning in a multi-task learning framework in order to make the best use of the noisy supervision signal. In particular, by sharing the encoder and decoder in various tasks, a more robust and effective encoder and decoder for the main photo-to-sketch synthesis task can be obtained. (2) Different from the existing unsupervised models based on cycle consistency (Figure 5.2(b)), our unsupervised learning tasks exploit the notion of shortcut cycle consistency: instead of passing through a different domain to get back to the input domain for reconstruction, our model takes a shortcut and completes a reconstruction within each domain. This is particularly effective given the large domain gap between photo and sketch.

Figure 5.1 shows that our model successfully translates photos to sketches stroke by stroke, demonstrating that the model has acquired an abstract and semantic understanding of visual objects. We compare against a number of state-of-the-art cross-domain image translation models, and show that superior performance is obtained by our model due to the proposed novel supervised and unsupervised hybrid learning framework with the shortcut cycle consistency. We also quantitatively validate the usefulness of the synthesised sketches for training a better fine-grained sketch-based image retrieval (FG-

SBIR) model.

Our contribution can be summarised as follows: (1) To our best knowledge, for the first time, the photo-to-sketch synthesis problem is addressed using a *learned* deep model, which enables stroke-level cross-domain visual understanding from a reference photo. (2) We identify the noisy supervision problem caused by subjective and varied human drawing styles, and propose a novel solution with hybrid supervised-unsupervised multi-task learning. The unsupervised learning is accomplished more effectively using a new shortcut cycle consistency constraint. (3) We exploit the synthesised sketches as an alternative to expensive photo-sketch pair annotation for training a FG-SBIR model. Promising results are obtained by using the synthesised photo-sketch pairs to augment manually collected pairs.

5.1 Methodology

5.1.1 Overview

We aim to learn a mapping function between the photo domain X and sketch domain Y, where we denote the empirical data distribution as $x \sim p_{data}(x)$ and $y \sim p_{data}(y)$ and represent each vector sketch segment as (s_{x_i}, s_{y_i}), a two-dimensional offset vector. Our model includes four mapping functions, learned using four subnets namely a photo encoder, a sketch encoder, a photo decoder, and a sketch decoder. They are denoted as E_p, E_s, D_p and D_s respectively.

5.1.2 Sub-Models

As illustrated by Figure 5.2(c), our model consists of four sub-models, each comprising an encoder subnet and a decoder subnet. (1) A supervised sub-model that translates a photo to a sketch; (2) a supervised sub-model that maps a sketch back to the photo domain;

(3) an unsupervised sub-model to reconstruct photo and (4) an unsupervised sub-model to reconstruct sketch. This means that our learning objective consists of two types of losses (to be detailed later): supervised translation loss for matching cross-domain and shortcut cycle consistency loss for traversing within domain.

5.1.3 Variational Encoders

The two encoders E_p and E_s are CNN and RNN respectively (see Figures 5.4(a) and (c)). In particular, E_s is a bidirectional LSTM. They take in either a photo or sketch as input and output a latent vector. They are variational because the latent vector is then projected into two vectors μ and σ with one fully connected (FC) layer. From the FC layer we construct our final embedding layer (bottleneck layer in each sub-model) by fusing it with a random vector, $\mathcal{N}(0, I)$, sampled from independent and identically distributed (IID) Gaussian distribution. To enable efficient posterior sampling, the reparameterisation trick is used as in [98]:

$$z = \mu + \sigma \odot \mathcal{N}(0, I) \tag{5.1}$$

5.1.4 Sketch Decoder

We build an LSTM-based sequence model as in [2] to sample output sketches segment by segment conditioned on the latent vector z (see Figure 5.4(b)). This is done by predicting each sketch segment offset $p(\Delta s_{x_i}, \Delta s_{y_i})$ using a Gaussian mixture model (GMM) and modelling pen state q_i for each time step as a categorical distribution, as detailed in [2]. To train the LSTM decoder, the reconstruction loss is formulated as:

$$\begin{aligned} &\mathcal{L}_{rnn}(S, \hat{S}) = \mathbb{E}_{x \sim S, y \sim \hat{S}} \\ &\Big[-\frac{1}{N_{max}} \Big(\sum_{i=1}^{N_s} \log(p(\Delta s_{x_i}, \Delta s_{y_i} | x, y)) - \sum_{i=1}^{N_{max}} \sum_{k=1}^{3} p_{k,i} log(q_{k,i} | x, y) \Big) \Big] \end{aligned} \tag{5.2}$$

where N_{max} represents the maximum number of segments in one sketch in the training set, and N_s denotes the actual length of segments for one particular sketch, thus N_s is usually smaller than N_{max}. Index i and k indicate the time step and one of three pen states, respectively.

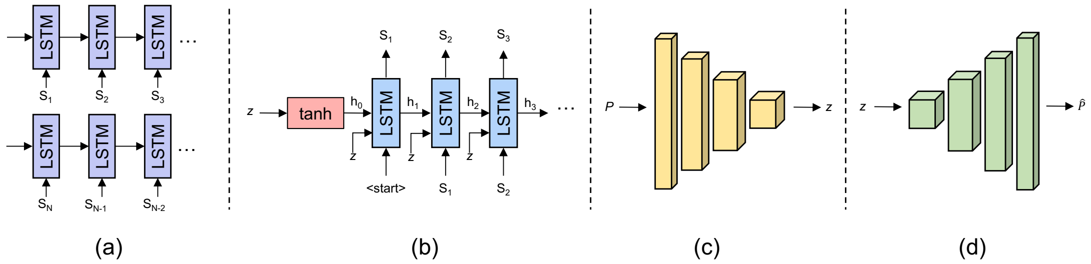


Figure 5.4: (a) bidirectional LSTM encoder E_s. (b) conditional LSTM decoder D_s. (c) generative CNN encoder E_p. (d) conditional CNN decoder D_p.

5.1.5 Photo Decoder

We use a CNN-based deconvolutional-upsampling block, as is commonly adopted by various generative tasks, where an l_2 loss

$$\mathcal{L}_{\rightarrow p}(P, \hat{P}) = \mathbb{E}_{x\sim P, y\sim \hat{P}}[||x - y||_2] \tag{5.3}$$

is used to measure the difference, which often leads to a blurry effect, known as the *regression to mean* problem [99]. An obvious solution is to add adversarial loss [26] for obtaining sharper photo visual effect. This was however not adopted because: (a) We did not observe improved photo-to-sketch synthesis, and even slightly worse due to the mode collapse issue, commonly observed with generative adversarial training [100]. (b) Synthesising photos is not the main goal of the model; it is used as an auxiliary task to help the main photo-to-sketch synthesis task. (c) Adding adversarial training loss to the synthesised sketch is not intuitive as the sketch is still vectorised and by doing the rasterisation, the gradient back-propagation will be stopped.

5.1.6 Shortcut Cycle Consistency

We might expect that learning a one-way mapping from photo to sketch should suffice, as paired examples exist for providing a supervision signal. However, as discussed, photo-sketch pairs provide a weak and noisy supervision signal, so such a one-way mapping function cannot be learned effectively. Our solution is to introduce two-way mapping using supervised learning and unsupervised reconstruction tasks. Since the four encoders and decoders are shared by these supervised and unsupervised tasks, they benefit from multi-task learning.

For the under-constrained mapping in the unsupervised self-reconstruction tasks, cycle consistency [6, 101] is developed to alleviate the *non-identifiable* [102] problem by reducing the space of possible mappings. This is achieved from the intuition that for each source image, the translation should be cycle consistent as to bring back to itself from the translated target domain. Taking photo to sketch translation for example, we have $x \rightarrow E_p(x) \rightarrow D_s(E_p(x)) \rightarrow E_s(D_s(E_p(x)) \rightarrow D_p(E_s(D_s(E_p(x)))$. However, since we do have noisy but paired data to provide weak supervision, the approximate posterior can actually be learned within each domain from the encoder's embedding. This is achieved by enforcing a variational bound and this is exactly where the shortcut can happen in the new cycle consistency proposed in this work.

Specifically, to form a photo to photo cycle now requires only traverse within domain, *i.e.*, $x \rightarrow E_p(x) \rightarrow D_p(E_p(x))$, which we term as shortcut cycle consistency. We find that apart from resulting in faster convergence in our supervised-unsupervised hybrid framework, our unsupervised sub-models with the shortcut cycle consistency can produce much better photo-to-sketch synthesis compared with the model learned with conventional cycle consistency. We postulate that given the large domain gap between photo and sketch, doing a long walk across domains potentially makes it harder to establish cross-domain correspondence. Formally, to enforce the shortcut cycle consistency, we minimise the following loss:

$$\begin{aligned}\mathcal{L}_{shortcut}(X,Y) &= \mathcal{L}_{\rightarrow s}(Y, D_s(E_s(Y))) \\ &+ \mathcal{L}_{\rightarrow p}(X, D_p(E_p(X)))\end{aligned} \tag{5.4}$$

5.1.7 Full Learning Objective

The four sub-models are learned jointly. Therefore, in additional to the unsupervised loss above, there are thus two supervised translation losses:

$$\begin{aligned}\mathcal{L}_{supervised}(X,Y) &= \mathcal{L}_{\rightarrow s}(Y, D_s(E_p(X))) \\ &+ \mathcal{L}_{\rightarrow p}(X, D_p(E_s(Y)))\end{aligned} \tag{5.5}$$

Furthermore, to enable efficient posterior sampling, we add KullbackLeibler (KL) losses for the bottleneck layer embedding space distributions to force the four sub-models to use a similar distribution to feed to their decoders. For simplicity, we combine them into one term:

$$\mathcal{L}_{KL} = \mathbb{E}_{x\sim X, y\sim Y, \hat{x}\sim\hat{X}, \hat{y}\sim\hat{Y}}[-\frac{1}{2}(1+\sigma^2-\exp(\sigma))|x,y,\hat{x},\hat{y}] \tag{5.6}$$

Our full objective thus becomes:

$$\mathcal{L}_{full}(X,Y) = L_{supervised}(X,Y) + \lambda_{shortcut}\mathcal{L}_{shortcut}(X,Y) + \lambda_{KL}\mathcal{L}_{KL} \tag{5.7}$$

where $\lambda_{shortcut}, \lambda_{KL}$ controls the relative importance of each loss. With the full loss, we aim to optimise:

$$\underset{E_p,E_s,D_p,D_s}{\operatorname{argmin}} \; L_{full}(X,Y) \tag{5.8}$$

5.2 Experiments

5.2.1 Datasets and Settings

5.2.1.1 Dataset Splits and Preprocessing

We use the publicly available QMUL-Shoe-Chair-V2 [61] dataset, the largest stroke-level paired sketch-photo dataset to date, to train and evaluate our deep photo-to-sketch synthesis model. There are 6,648 sketches and 2,000 photos for the shoe category, where we use 5,982 and 1,800 of which respectively for training and the rest for testing. For chairs, we split the dataset as following strategy: 300/100 photos, 1275/725 sketches for training/testing respectively. It is guaranteed that each photo is paired with at least one human sketch. We scale and centre crop the photos to 224 × 224 pixels and pre-process original sketches via stroke removal and spatial sampling to reduce to number of segments to the level suitable for LSTM-based modelling.

5.2.1.2 Pretraining on QuickDraw Dataset

Due to the limited number of sketch-photo pairs in QMUL-Shoe-Chair-V2, we pretrain our model with 70,000 shoe and 70,000 chair training sketches from the QuickDraw dataset [2]. Despite the fact that only abstract iconic vector sketches exist with no associated photos, we form our pretrained photos by transforming sketches to raster pixel images.

5.2.1.3 Implementation Details

Our CNN-based encoder and decoder, E_p and D_p consist of five stride-2 convolution layers, two fully connected layers and five fractionally-strided convolutional layers with stride 1/2, similar to [4] but without skip connections. We use instance normalisation instead of batch normalisation as in [103]. We adopt bidirectional and unidirectional

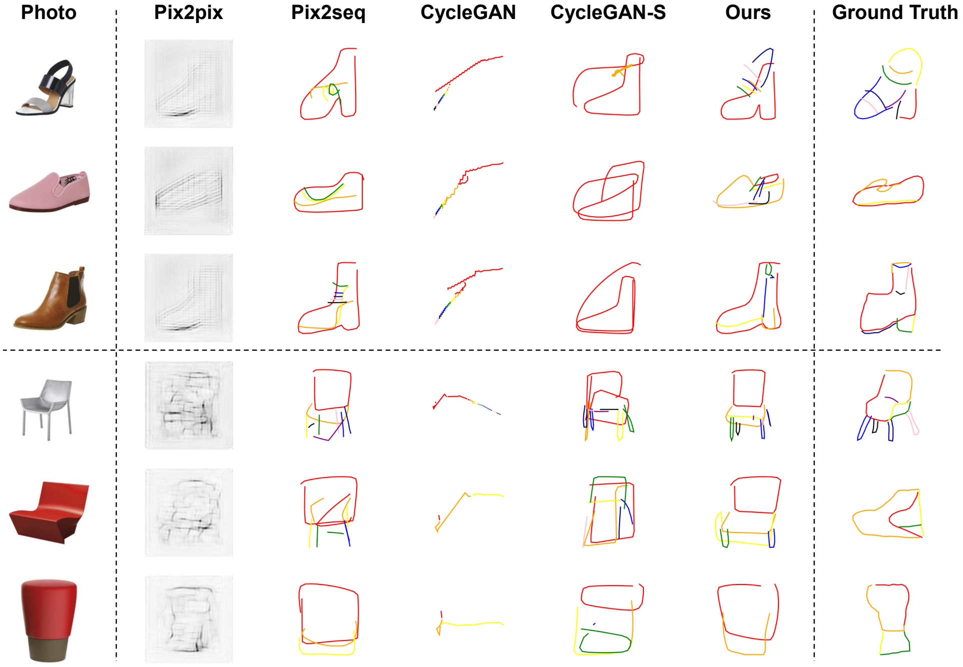


Figure 5.5: Photo-to-sketch synthesis on the QMUL-Shoe-Chair-V2 test splits. From left to right: input photo, Pix2pix [4], Pix2seq [5], CycleGAN [6], CycleGAN with supervised translation loss, ours and ground truth human sketch.

LSTM for our RNN encoder E_s and decoder D_s respectively, while keeping other learning strategies the same as [2]. We implement our model end-to-end on Tensorflow [104] with a single Titan X GPU. We set the importance weights $\lambda_{shortcut} = 1$ and $\lambda_{KL} = 0.01$ during training and find this simple strategy works well. Both pretraining and fine-tuning stages are trained for a fixed 200,000 iterations with a batch size of 100. The model is trained end to end using the Adam optimiser [105] with the parameters $\beta_1 = 0.5, \beta_2 = 0.9, \epsilon = 10^{-8}$. A fixed learning rate of 0.0001 is adopted for experiments.

5.2.2 Evaluation Metric

Evaluating the quality of synthesised images is still an open problem. Traditional maximum likelihood approaches (*e.g.*, kernel density estimation) fail to offer a true reflection of the synthesis quality, as validated in [106]. Consequently, most recent studies either run human perceptual studies by crowd-sourcing or explore computational metrics attempting to predict human perceptual similarity judgements [107]. Our measures fall into the latter by discriminatively answering two questions: (i) How recognisable can the synthesised sketch be when evaluated with a recognition model trained on human sketch data? (ii) How realistic and diverse are the synthesised sketches, so that they can be used as queries to retrieve photos using a FG-SBIR model trained on photo-human sketch pairs? A good score under these metrics requires synthesised sketches to be both realistic and instance-level identifiable. The metric thus shares the same intuition behind the "inception score" [100]. More specifically, the two metrics are: (1) **Recognition-Accuracy**: We feed the synthesised sketches into the Sketch-a-Net [31] model, which is trained to recognise 250 real-world sketch categories with super-human performance. The assumption is that if a synthesised sketch can be recognised correctly as the same category as the corresponding photo, we can conclude with some confidence that it is category-level realistic. (2) **FG-SBIR Retrieval-Accuracy**: Since our data are from the same category (either shoe or chair), the recognition-score could still be high if the model learns to one specific object instance regardless of the input photo instances (*i.e.*, the typical symptom of mode collapsing [100]), or if the synthesised sketches are diverse but hardly resemble the object instances in the corresponding photos. To overcome this problem, the FG-SBIR accuracy is introduced as a harder metric. We retrain the model of [3] on the QMUL-Shoe-Chair-V2 training split [61] and used the synthesised sketches to retrieve photos on the test-split.

5.2.3 Competitors

For fair comparison, we implement all the competitors under the same architecture and training strategies as our model, except for CycleGAN [6], where we have to add two discriminators for adversarial training to compensate for its unsupervised setting. **Pix2pix** [4]: We compare with replacing vector sketch images with raster sketch images, where translation happens within the pixel space. We tried different state-of-the-art cross-domain translation models [4, 108, 109], but did not find much difference between them.

We thus only report the results of the model in [4] as a representative one. **Pix2seq** [5]: This corresponds to the ablated version of our full model: a one-way photo-to-sketch supervised translation model with vector sketch as output. This is similar to [5], which was originally designed for better sketch reconstruction, now re-designed and re-purposed for the photo-to-sketch translation task. **CycleGAN** [6]: This is proposed to specifically target image-to-image translation with the absence of paired training examples. Cycle consistency is enforced to alleviate the highly under-constrained setting of the problem. **CycleGAN-Supervised (CycleGAN-S)**: Additional supervised learning modules (two discriminators for adversarial training) are added on top of CycleGAN to give a level playing field. This can be considered as an alternative supervised-unsupervised hybrid model.

5.2.4 Qualitative Results

As illustrated in Fig. 5.5, all four competitors fail to generate high quality sketches that match with the corresponding photo. Our model, in contrast, is able to sketch object abstractly but semantically. Interestingly, our model produces some sketches with certain level of fine-grained details, which is extremely challenging given the highly noisy supervision signals as shown in Fig. 5.3. In some cases, *e.g.*, the third row shoe example, the synthesised sketch matches the actually object shape better and contains more fine-grained details compared to the human sketch.

The competitors suffer from various problems. We observe complete model collapse when using CycleGAN under unsupervised setting, which suggests that CycleGAN may only works with unpaired training examples under a strong cross-domain pixel-level alignment assumption. After injecting supervision into CycleGAN (CycleGAN-S), the synthesised results get better but still suffers from regular noisy stroke generation, *i.e.*, it seems that a random meaningless stroke is always sketched on a shoe. In contrast, our model with shortcut cycle consistency does not suffer from such issue. This is because our model takes a shortcut from the bottleneck, which eases the burden on optimisation and

Method	ShoeV2				ChairV2			
	Recognition		Retrieval		Recognition		Retrieval	
	Top 1	Top 10	Top 1	Top 10	Top 1	Top 10	Top 1	Top 10
Human sketch [61]	36.50%	70.00%	30.33%	76.28%	10.00%	35.00%	47.68%	89.47%
Pix2pix [4]	0.00%	0.00%	0.50%	7.50%	0.00%	0.00%	2.00%	16.00%
Pix2seq [5]	51.50%	86.00%	4.50%	26.00%	5.00%	51.00%	3.00%	31.00%
CycleGAN [6]	0.00%	0.00%	0.50%	4.00%	0.00%	8.00%	1.00%	7.00%
CycleGAN-S	18.00%	51.50%	2.00%	18.00%	12.00%	**55.00%**	6.00%	33.00%
Our full model	**53.50%**	**90.00%**	**6.00%**	**28.50%**	**13.00%**	**55.00%**	**8.00%**	**36.00%**

Table 5-A: Recognition and retrieval results obtained using the synthesised sketched. Numbers in red and blue indicate the best and second-best performance among compared models. The results are in top-1 and top-10 accuracy.

enhances the representation power of the encoder. We also witness some success using the Pix2seq model – the sketch looks adequate on its own, but when compared with the corresponding photo, it does not bear much resemblance, often containing some wrong fine-grained details, *e.g.*, ankle strap of the first-row shoe. This supports our hypothesis that one-way image-to-image translation is not enough to deal with the highly-noisy paired training data. Finally, the worst results are obtained by the Pix2pix model which is the only model that treats sketch as a raster pixel image. The synthesised sketches are blurry and lack sharp and clean edges. This is likely caused by the fact that the model pays too much attention to handling the empty background which is also part of data to model with the raster image format.

5.2.5 Quantitative Results

We compare the performance of different models evaluated using the two metrics (Sec. 5.2.2) in Table 5-A. The following observations can be made: (i) Under the recognition metric, our model beats all the competitors. Interestingly it also beats human, showing our superior category-level generative realism. (ii) Under the retrieval metric, our model still outperforms all competitors on both datasets. However, this time, the gap to the human sketches' performance is big. This suggests that when humans draw a sketch of a specific object given a reference photo, attention is paid mainly to fine-grained details for distinguishing different instances, rather than the category-level realism. Never-

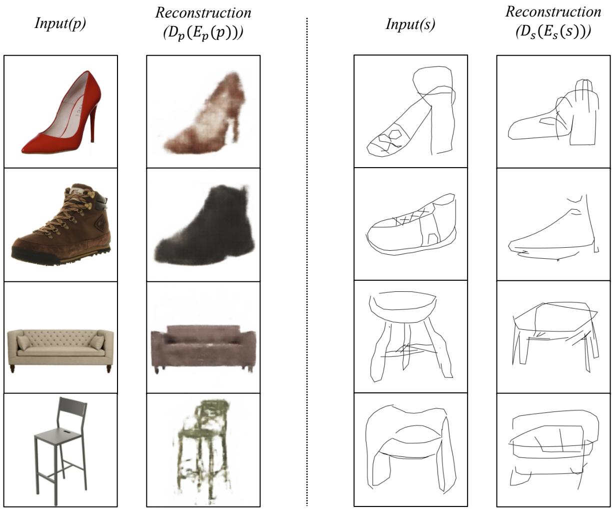

Input(p)
Reconstruction $(D_p(E_p(p)))$
Input(s)
Reconstruction $(D_s(E_s(s)))$

Figure 5.6: Sketch-to-sketch and photo-to-photo reconstruction results on QMUL-Shoe-Chair-V2 dataset.

theless, compared to the chance level (0.5% Top 1 for ShoeV2 and 1% for ChairV2), our model's performance suggests the synthesised sketches do capture some instance-identifiable details. (iii) The strongest competitor on ShoeV2 is Pix2seq [5]. However, its place is taken by CycleGAN-S on ChairV2. This is expected: the ChairV2 dataset is much smaller than ShoeV2, posing difficulties for a pure supervised-learning based approach. The unsupervised CycleGAN yields poor performance all the time due to model collapse, but its supervised learning boosted version CycleGAN-S fares quite well on the small ChairV2 dataset. This further validates our claims that a hybrid model is required and our shortcut consistency is more effective than the full cycle consistency. In summary, our model quantitatively beats all competitors under both metrics and even shows better category-level realism than human sketches, in accordance with qualitative observations.

5.2.6 Image Reconstruction Quality

In Figure 5.6, we show a few samples of the reconstruction results obtained using our unsupervised sub-models, *i.e.*, $sketch \rightarrow D_s(E_s(sketch))$ and $photo \rightarrow D_p(E_p(photo))$, with our shortcut cycle consistency. We observe that the reconstructed photos are quite

Dataset	Top 1	Top 10
Without pretraining on synthetic data	30.33%	76.28%
With pretraining on synthetic data	32.43%	77.48%

Table 5-B: Evaluation of the contribution of synthetic sketch pretraining on FG-SBIR.

close to the input, despite the expected blurry effects (as explained in Sec. 5.1.1). For sketches, due to the existence of the KL loss (Eq. 5.6), the RNN-based decoder suffers significant reconstruction degradation, which is also shown in [5]. However, as mentioned earlier, among the four sub-models, only the photo-to-sketch one is what we are after, and the other three are designed as auxiliary tasks to learn a better encoder and decoder to serve the main sub-model. In this case, it appears that the $sketch \to D_s(E_s(sketch))$ sub-model has sacrificed its own performance to help the main sub-model.

5.2.7 Data Augmentation for FG-SBIR

In this experiment, we evaluate whether the synthesised sketches using our model can be used to form some additional photo-sketch pairs to train a better FG-SBIR model. More concretely, we collect 1800 photos from a different shopping website (Selfridge's), called ShoeSF, which have no overlap with the ShoeV2 photos. Figure 5.7 shows some examples of newly collected dataset, and the synthesised sketches based on out model. We then apply our model trained on ShoeV2 to generate sketches for ShoeSF to form some additional photo-sketch pairs. They are then used to pretrain the FG-SBIR model in [3] before fine-tuning on the ShoeV2 provided photo-sketch pairs. Table 5-B shows that using the synthesised data can boost the performance by 2.10% Top 1. One possible reason for the limited improvement can be the synthesis quality still needs to be improved.

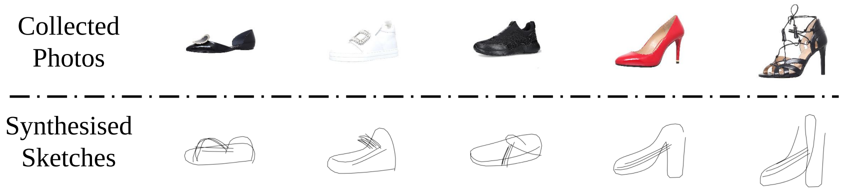


Figure 5.7: Examples of our newly collected dataset and the corresponding synthesised sketches.

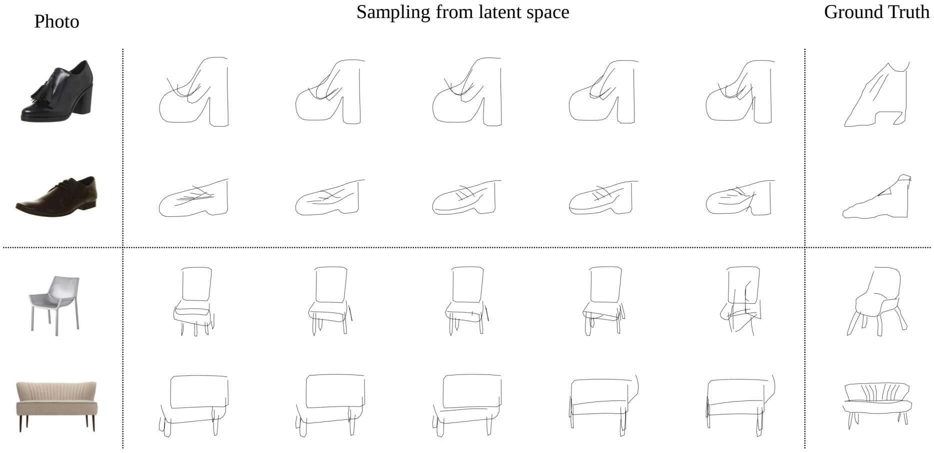


Figure 5.8: Exploring the embedding space by interpolating the latent vector.

5.2.8 Exploring the Embedding Space

With the help of latent vector z and the KL loss, we are able to explore the embedding space from CNN encoder. We sample the latent vector by interpolate the random noise from -1 to 1 with stride 0.5, and then visualise the synthesised sketches as Fig. 5.8.

5.3 Summary

We proposed the first deep stroke-level photo-to-sketch synthesis model that enables abstract stroke-level visual understanding of an object in a photo. To cope with the noisy supervision of photo-human sketch pairs, we proposed a novel supervised-unsupervised hybrid model with shortcut cycle consistency. We show that our model achieves superior performance both qualitatively and quantitatively over a number of state-of-the-art alternatives. We also applied our synthetic sketches as a mean of data augmentation for the FG-SBIR task, obtaining promising results. This application indicating that the synthesis task can help retrieval task by enriching the training data, but the limitations are the application may be constrained by the qualities of synthesised sketches.

Chapter 6

Generalisation for FG-SBIR

Sketch modality is considered with promising advantages like informative, conveying semantic meaning, encoding temporal information and so on. However, sketch collection and annotation are costly both in time, labour and price, thus will bring trouble to the sketch researches. The high expense of data expense drives us to consider the generalisation learning of the sketch related models. In this chapter, we will discuss how to design a generalised fine-grained SBIR model, which will be able to generalise well on novel categories of sketches, without asking volunteers to label a large amount of sketches in high expense.

The problem of fine-grained sketch-based image retrieval (FG-SBIR) has been studied intensively. A FG-SBIR model is used to retrieve the most similar photo according to the query sketch. To search for the target photo of the same object given the sketch, most recent FG-SBIR models apply deep convolutional neural networks (CNNs) to learn a feature embedding space where sketch modality and photo modality are aligned based on feature distances [3, 14, 23]. These models are normally trained and tested on the same category. In practice, however, the object may come from very different categories for the training data, thus the model may fail to provide a satisfying retrieval result on the new category as the model is not be able to provide a generalised performance. Deep FG-SBIR models that are directly applied to new category without model updating are known to suffer from performance degradation [3, 14, 23], thus suggesting model overfitting and poor domain generalisation.

In this chapter, we aim to learn a generalisable FG-SBIR model. Such a model is trained on a set of source categories, and should be able to generalise to any new unseen category for effective FG-SBIR without any model updating. Such a model thus needs to solve a generalisation problem with different class label spaces for different datasets/domains. A generalisable FG-SBIR model has great value for real-world large-scale deployment. Specifically, when a customer purchases a FG-SBIR system for a specific retrieval network, the system is expected to work out-of-the-box, without the need to go through the tedious process of data collection, annotation and model updating/fine-tuning.

Surprisingly, there is very little prior study of this topic. Existing FG-SBIR works occasionally evaluate their models' cross-category generalisation, but no specific design is made to make the models more generalisable. Beyond FG-SBIR, the problem of domain generalisation (DG) has been investigated in deep learning, with some recent few-shot meta-learning approaches also adapted for domain generalisation. However, existing domain generalisation methods [52, 53, 55, 56] assume that the source and target domain have the same label space; whilst existing meta-learning models [57, 60, 110, 110, 111] assume a fixed number of classes for target domains and are trained specifically for that number using source data. They thus have limited efficacy for FG-SBIR, where target domains have a different and variable number of categories and instances.

Our solution to generalisable FG-SBIR is based on a novel Domain-Invariant Mapping Network(DIMN). DIMN is designed to learn a mapping between a photo image and its instance classifier weight vector, *i.e.*, it produces a classifier using a single shot. Once learned, for a target domain, each photo will be fed into the network to generate the weight vector of a specific linear classifier for the corresponding instance. A query sketch

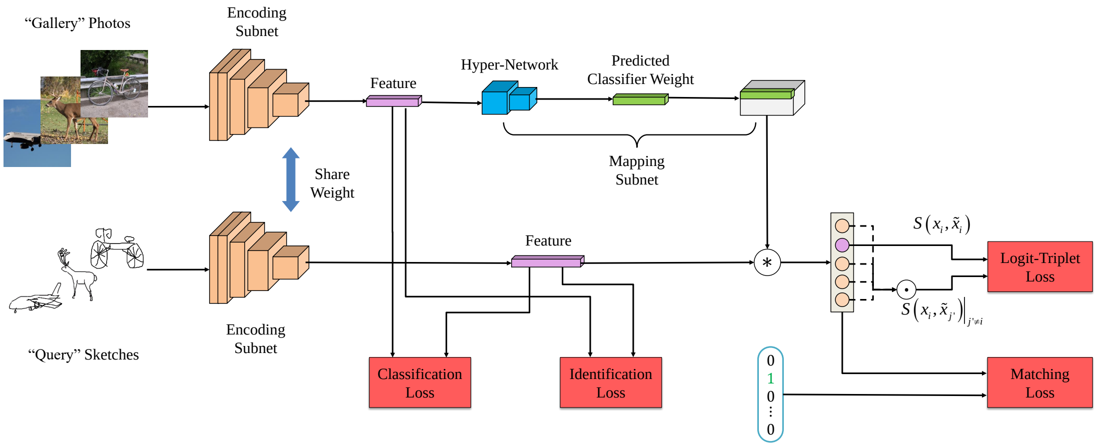


Figure 6.1: The proposed Domain-Invariant Mapping Networkfor generalised FG-SBIR.

will then be matched with the gallery photo using the classifier by computing a simple dot product between the weight vector and a deep feature vector extracted from the query sketch. To make the model generalisable to different categories, we follow a meta-learning pipeline and sample a subset of source training tasks (categories) during each training episode. However, the model is significantly different from conventional meta-learning methods in that: (1) No model updating is required for the target categories. (2) Once trained, the model can be used to match an arbitrary number of categories in the target application scenario.

Our contributions are as follows: (i) For the first time, the generalisation problem in FG-SBIR is explicitly highlighted and also tackled by designing a FG-SBIR model that is tailor-made for coping with unknown target categories. (ii) A novel Domain-Invariant Mapping Network(DIMN) is proposed whose generalisability comes from its ability to map an image directly into an instance classifier. Extensive experiments validate the generalisability of our DIMN and suggest that it is superior to the baseline methods.

6.1 Methodology

6.1.1 Overview

We study a generalised fine-grained sketch-based image retrieval problem, where in the training stage, we have the access to M categories, $\mathcal{D}_1$, $\mathcal{D}_2$, ... and $\mathcal{D}_M$, and each category has its own instance label space (indicating whether the sketch and photo coming from the same instance or not). The trained model will be deployed directly to a new category, and is expected to work without any further model update. To this end, we propose a Domain-Invariant Mapping Network(DIMN), illustrated in Fig. 6.1. The training images are organised into gallery and probe sets to simulate the testing scenario where a query sketch is compared against a gallery set for matching. The proposed network consists of three modules: (1) Two weight-tied base networks, the encoding subnets, which serve as feature extractors for probe sketches and query photos respectively. (2) A hyper-network [112], namely mapping subnet, which takes the gallery photo embedding as input and tunes it into the instance classifier's weight vector that represents the instance identity of the gallery photo. (3) A logit-triplet ranking loss to better predict the matching relations. We will detail the design of each module in the following sections.

6.1.2 Encoding Subnet

For the encoding subnet, we use MobilenetV2 [113] – a lightweight CNN with competitive performance compared to heavier alternatives such as ResNet [114] and InceptionV3 [115]. We found it to be both more efficient and more effective for our generalised FG-SBIR task.

As shown in Fig. 6.1, the two Siamese encoding subnets in DIMN are used in the gallery and probe branches respectively. To generate the inputs for both branches, we follow a specific mini-batch sampling procedure. Assuming we have C unique instances in total in the overall M training categories, and more specifically, C_i is the number of instances for ith category. We sample $M_b (M_b \ll M)$ categories from all the categories, and for each category, sample C_b $(C_b < C_i, i \in 1...M)$ instances randomly for each mini-batch. For each instance y_i, we further sample one sketch and one photo, of which we assign the photo as gallery p_i and the sketch as probe s_i. Therefore, we have $M_b C_b$ sketch/photo pairs in a mini-batch, as illustrated in Fig. 6.2.

Assuming the encoding subnet produces a D-dimensional feature vector, the first training objective for DIMN is a classification loss for the total M categories, denoted as $\mathcal{L}_{\text{cat}}$,

$$\mathcal{L}_{\text{cat}} = \sum_{i=1}^{C_b} \text{Cross\_Entropy}(l_i, \text{Softmax}(f_\theta(g_\phi(s_i)))) \tag{6.1}$$

where s_i is the query sketch and l_i is the one-hot encoding of its category label (a M-dimensional unit vector). $g_\phi(\cdot)$ is the encoding subnet parameterized by ϕ. $f_\theta(\cdot)$ is the category classifier parameterized by θ where $\theta \in \mathbb{R}^{D \times M}$. Operation Cross_Entropy and Softmax are described in and , respectively.

$$\text{Softmax}(x) = \frac{e^{x_i}}{\sum e^{x_i}} \tag{6.2}$$

$$\text{Cross\_Entropy}(t, y) = \sum t_i \text{log} y_i \tag{6.3}$$

Notice that here we use cross-entropy loss as the classification loss. Admitedly, alternative losses can also be used for classification, like regression loss and KullbackLeibler (KL) divergence losses. However, the former is sensitive to outliers while the latter is exactly the same in the term of optimisation.

To increase the discriminativity of the learned feature, we also have a second training objective for DIMN , which is an identification loss for the total C instances, denoted as

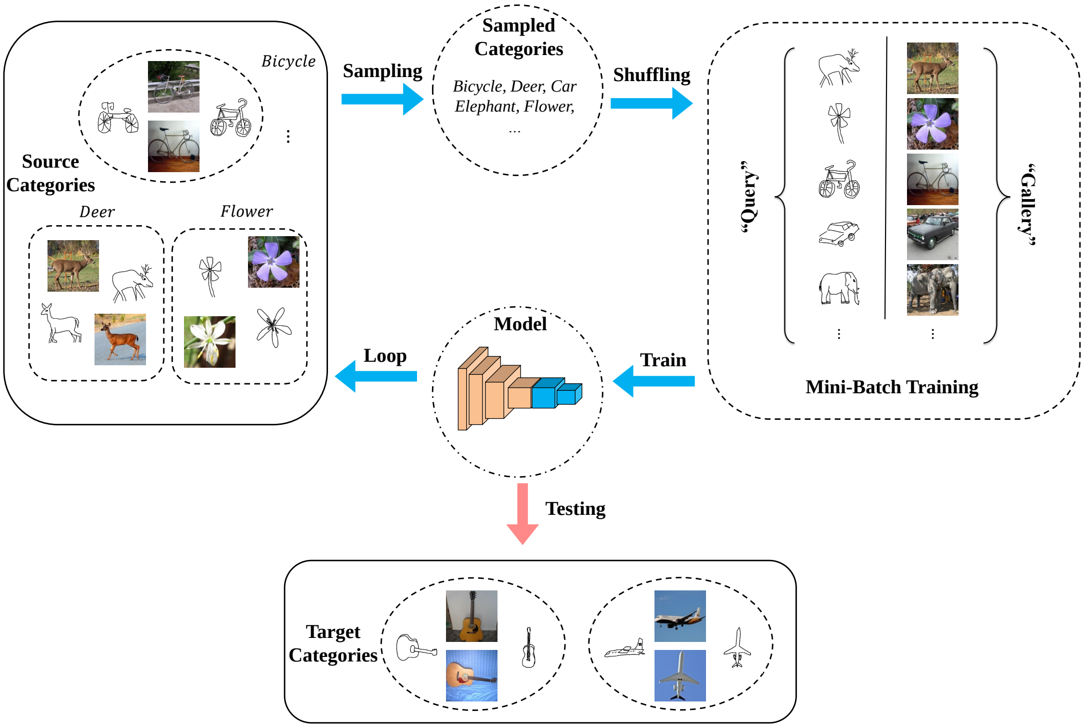


Figure 6.2: An illustration of the mini-batch sampling strategy.

$\mathcal{L}_{\text{id}}$,

$$\mathcal{L}_{\text{id}} = \sum_{i=1}^{C_b} \text{Cross\_Entropy}(t_i, \text{Softmax}(f'_\theta(g_\phi(s_i)))) \tag{6.4}$$

Similarly, t_i is the one-hot encoding of its instance label (a C-dimensional unit vector). $g_\phi(\cdot)$ is the encoding subnet parameterized by ϕ. $f'_\theta(\cdot)$ is the instance classifier parameterized by θ' where $\theta' \in \mathbb{R}^{D\times C}$.

6.1.3 Mapping Subnet

Conventional identification loss like $\mathcal{L}_{\text{id}}$ can help discriminative representation learning but it also suffers from overfitting problem, especially when it is applied to an unseen domain. To alleviate this problem, we further designed a domain-independent classification module to predict the matching relations in the training domains. Specifically, we propose to use a dynamical model, named as the "mapping net" to dynamically generate the classifier weights for the matching network. The matching network is then using this synthesised weights to classify the matching relations, instead of the static learned model parameters.

The deep feature vector, extracted from each gallery photo using the encoding subnet, is then fed into a mapping subnet to compute a classifier weight vector for the corresponding instance. Formally, given an instance of the jth category from the gallery photo branch, denoted as p_j. Instead of learning the jth classifier weight vector $\theta_{\cdot,j}$ as part of the model parameters, as in a conventional classification CNN, we *generate* it as a layer of the network using p_j as input. We thus have:

$$\hat{\theta}_{\cdot,j} = h_\omega(g_\phi(p_j)), \tag{6.5}$$

where the mapping subnet $h_\omega(\cdot)$ can be understood as a hyper-network [112] since it generates the parameters for another neural network (the probe sketch branch). Here we simply apply a multi-layer perceptron (MLP) as the basic architecture of our mapping subnet. Note that we omit the bias term in the weight generation for simplicity.

Given a gallery photo p_j, and a query sketch s_i, the mapping subnet generates an identity classifier weight vector $\hat{\theta}_{\cdot,j}$ based on the gallery photo, p_j. We then take the dot product of the generated classifier weight vector $\hat{\theta}_{\cdot,j}$ and the query sketch feature $g_\phi(s_i)$, to produce a logit vector p whose elements corresponding the identity of p_j: $p_j = h_\omega(g_\phi(p_j)) \cdot g_\phi(s_i)$. Passing the vector p into a softmax layer then gives us the predicted probability of how likely the input instance s_i in the query sketch branch is matched with the instance p_j in the gallery photo branch. The ground truth label y for the matching network will be 1 if s_j matches with p_j, and 0 otherwise. y can then be used for computing a classification loss. The new classification loss the matching network is named as matching loss, as described as below,

$$\mathcal{L}_{\text{mat}} = \sum_{i=1}^{C_b} \text{Cross\_Entropy}(y_i, \text{Softmax}(\hat{W}_b^T g_\phi(x_i))) \tag{6.6}$$

where $[x_1, x_2, \ldots, x_{C_b}]$ are probe branch inputs. $\hat{W}_b$ is the classifier weights produced by the hyper network and parametarised by the gallery photos, as shown in Eq.6.7.

$$\hat{W}_b \leftarrow \hat{\theta}_{\cdot,j} \quad \forall j \in [1, 2, \ldots, C_b], \tag{6.7}$$

Note that the logit vector p is a C-dimensional vector which can be of very high-dimensionality with a large number of instances in the source domains. If we follow the standard meta-learning practice and reduce the dimensionality to the much smaller number M_bC_b, the model training becomes tractable. However, we then lose the discriminative power: the mapping network is trained to perform a much easier task of classifying C_b people rather than C. Thus we then designed a logit-triplet loss to keep the discriminativity power in compensation.

6.1.4 Logit-triplet Loss

We further introduce a specific triplet loss built on our matching network, named as logit-triplet loss. As a by-product of building the mapping subset, for every instance in the query sketch branch, s_i, we can find its only positive pair p_i in the gallery photo branch and compute the logit: $p = h_\omega(g_\phi(p_i)) \cdot g_\phi(s_i)$, meanwhile, we can also find negative pairs by computing: $n = h_\omega(g_\phi(p_j)) \cdot g_\phi(s_i)|_{j \neq i}$ among all the gallery photos. Both p and n will be further normalised to produce valid probabilities as the result of applying softmax function in Eq. 6.6. Denote the normalised p and n as $S(s_i, p_i)$ and $S(s_i, p_{j'})|_{j' \neq i}$, respectively, which also means the similarity score or matching probability between the query sketch and gallery photo pairs. We can then adopt the following logit-triplet loss with the hard mining [116],

Algorithm 1 Training Domain-Invariant Mapping Network

```
Input: $\mathcal{D}_1$, $\mathcal{D}_2$, ... and $\mathcal{D}_M$;
for $t = 1$ to Max_Iter do
    Sample a subset of categories $\mathcal{D}_l \in \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_M\}$
    Sample $\{(s_1, p_1, y_1), \dots, (s_{C_b}, p_{C_b}, y_{C_b})\} \in \mathcal{D}_l$
    $\hat{\theta} \leftarrow h_\omega(g_\phi(p))$
    Calculate losses: $\mathcal{L}_{\text{id}}$, $\mathcal{L}_{\text{mat}}$, and $\mathcal{L}_{\text{tri}}$
    Optimise $\mathcal{L}_{\text{full}}$ via the optimiser
end for
```

$$\mathcal{L}_{\text{tri}} = \sum_{i=1}^{C_b} \max\left(0, \Delta + \max_{j' \neq i} S(s_i, p_{j'}) - S(s_i, p_i)\right) \tag{6.8}$$

Note that there are several main differences between our logit-triplet loss and the traditional triplet loss: (i) we are using a similarity scalar of the matching probability rather than the Euclidean distance between features. Therefore, we are actually hoping this similarity scalar between the anchor and positive pairs larger, while the scalar between the anchor and the negative pairs smaller. (2) The scalar in the logit-triplet loss is normalised by the softmax layer, rather than the normally used l_2 normalisation.

6.1.5 Training Objective

The model is trained in an end-to-end fashion and the full training objective $\mathcal{L}_{\text{full}}$ is a weighted sum of Eq. 6.4, Eq. 6.6, and Eq. 6.8.

$$\mathcal{L}_{\text{full}} = \mathcal{L}_{\text{id}} + \lambda_1 \mathcal{L}_{\text{mat}} + \lambda_2 \mathcal{L}_{\text{tri}} \tag{6.9}$$

The training pipeline is summarised in Alg. 1.

6.1.6 Model Testing

Trained in a meta-learning pipeline by sampling training categories in each episode, both the encoding subnet ($g_\phi(\cdot)$) and mapping subnet ($h_\omega(\cdot)$) in our DIMN are supposed to be category invariant. During the testing stage, given a query sketch s_i, and a gallery photo p_j, we directly take the logits (or probability after the softmax layer) $h_\omega(g_\phi(p_j)) \cdot g_\phi(s_i)$ as the ranking score. It is importantly to point out that: (1) Although it looks like a one-shot learning method, DIMN is a domain generalisation method as the model itself (*i.e.*, $g_\phi(\cdot)$ and $h_\omega(\cdot)$) is fixed once trained on training source. (2) Conventional deep FG-SBIR models only have an encoding network $g_\phi(\cdot)$, and uses the Euclidean distance between $g_\phi(s_i)$ and $g_\phi(p_j)$ as the ranking score. Comparing to them, our DIMN has very similar inference cost during the testing stage.

6.2 Experiments

6.2.1 Dataset and Settings

6.2.1.1 Generalised FG-SBIR benchmark

We introduce a generalised FG-SBIR benchmark to evaluate the generalisation ability of a FG-SBIR model. We aim to simulate a real-world scenario where a FG-SBIR model is trained with several seen categories, and most likely exposed to an unseen category during the inference stage. To this end, we use the Sketchy dataset [14], which is the large-scale FG-SBIR dataset covering 125 categories. We deliberately select 21 categories to form the target unseen categories, and leave the rest 104 categories as the source training categories. When selecting the unseen categories, we follow the principle of testing categories are excluded in the ImageNet 1000 categories [117], thus ensuring that these categories are also unseen to our ImageNet pretrained model, similar to [118]. All the images in the Sketchy dataset, still keeps the original train/test splits for each

Data Statistics	Train		Test	
	Sketch	Photo	Sketch	Photo
# Categories	104	104	21	21
# Instances	9360	9360	210	210
# Images	54228	9360	1069	210

Table 6-A: Dataset statistics of our generalised FG-SBIR benchmark based on Sketchy.

category, *i.e.*, we only use the train split from the seen categories, and test on the test split of the unseen category. The reason is that for the future works we are able to use the extra split to further tune our model. Another reason is that we can evaluate the model performance on the standard testing split only, so our testing gallery size is the same for each category. The dataset details are listed in Table 6-A. Note that the total number of training categories is $C = 104$ with 54228 training sketches and 9360 training images, while for testing, we have 21 unseen categories with 1069 testing sketches and 210 testing images. The gallery size for the testing stage is thus 210.

6.2.1.2 Implementation Details

We use MobileNetV2 [113] as the encoding subnet, with width multiplier of 1.4. The output feature dimension is thus $1,792$. Our mapping subnet is composed of three fully-connected (FC) layers. The output size is set to the same as the input size, as the dimension of the classifier weights should be the same as the feature dimension. The dimensions for the three FC layers are all set to $1,792$. The logit-triplet loss margin (Eq. 6.8) is set to $\Delta = 0.8$. The weights for the classification loss and logit-triplet loss are set as equal, *i.e.*, $\lambda_1 = \lambda_2 = 1$. We implement our model in Tensorflow [104] and train it with a single Titan X GPU. The model is trained for a fixed $180,000$ iterations with batch size 64, which means in each iteration, we sample $2(M_b)$ categories in each batch, and for each categories, we sample $8(C_b)$ sketch and photo pairs; each comes with 8 sketches used as the "probe" while the 8 photos are used as the "gallery". Exponential decay learning rate scheduling is used with initial rate 0.00035 and ending with 0.0001.

Adam optimiser [105] is used for all experiments. For the training variables in the network, we use Xavier initialisation [119] for a more robust performance.

6.2.1.3 Evaluation Metrics

We follow the standard single-view evaluation protocols on the testing dataset like most existing FG-SBIR methods [3, 14]. Two commonly used evaluation metrics are used. The first is cumulative matching characteristics (CMC). We report the CMC at rank-k, where $k = 1, 10$, representing the ranking accuracy of the target identities in the top k results. The second metric is the mean average precision (mAP), which reflects the overall ranking quality rather than looking at top k positions only. In summary, we report the mAP and Top 1 and Top 10 accuracy on the test set during the evaluation.

6.2.2 Comparisons against State-of-the-art

6.2.2.1 Baselines

We compare our model with the state-of-the-art baseline [3]. Besides this baseline, we also consider the verification based ranking network [120] and one naive baseline where only category-level information is used, termed as Ver and Clf, respectively. In the [3] baseline (termed as Tri), we train a triplet model which is effective for the fine-grained ranking. We uses the same MobileNetV2 backbone for fair comparison to our DIMN. In addition, we also designed a baseline in a popular way for ReID task [87], which engaged an identification loss, termed as IDE. The meta-learning method, PPA [60] is effective for alleviating over-fitting in few-shot learning-to-learn, and can be adapted for the generalised FG-SBIR problem here (unlike most others that require model updating). Notice that for all the baselines except PPA, category-level classification loss is added. PPA model do not need the extra classification loss as it starts from the feature and the feature actually comes from a category-level classification model. Note that the our

Sketchy Dataset	Top 1	Top 10	mAP
Clf	12.63%	49.67%	23.86%
Ver [120]	30.59%	80.45%	46.70%
Tri [3]	36.20%	76.05%	50.05%
IDE [87]	37.23%	85.50%	52.81%
PPA [60]	36.86%	85.03%	52.31%
Ours	**39.01%**	**86.44%**	**53.81%**

Table 6-B: Comparison against state-of-the-art methods.

prior methods described in Chapter 3 and 4 cannot apply here, as no attribute or text information are available.

6.2.2.2 Results

We compare the proposed method with several baselines on the target unseen dataset. The retrieval performance is listed in Table 6-B. The following observations can be made: (1) Overall, our method achieves the best result on this generalised FG-SBIR task among all compared methods. (2) The triplet baseline indeed is very strong in this task, given the second best performance among all the methods, but can not generalise as good as our method which has the generalisation module specifically designed.

6.2.3 Generalised performance for conventional SBIR task

Our model is also capable to the conventional SBIR task, as it seeks the semantic understanding of the sketch and photo content, with the help of classification loss. We therefore also evaluate the generalised performance of the proposed model and baseline methods under the conventional SBIR setting. We take the same trained models without any further model updating and using ranking accuracy (Top 1 and Top 10) and mAP as the evaluation metric and report the generalised SBIR performance. The result is shown in Tab. 6-C. From Tab. 6-C, we can see that our method is able to achieve the best generalised SBIR performance overall. Admittedly, the IDE baseline can get a higher mAP, but our method can get a more satisfied Top 1 accuracy, which is more beneficial espe-

Sketchy Dataset	Top 1	Top 10	mAP
Clf	53.23%	92.52%	39.49%
Ver [120]	69.50%	97.75%	43.86%
Tri [3]	68.57%	96.63%	35.13%
IDE [87]	78.77%	98.06%	**55.75**%
PPA [60]	78.58%	98.32%	50.54%
Ours	**79.70**%	**99.06**%	50.81%

Table 6-C: Generalised performance under conventional SBIR setting.

Sketchy Dataset	Top 1	Top 10	mAP
Supervised [3]	45.74%	89.80%	60.80%
Ours	39.01%	86.44%	53.81%

Table 6-D: Comparison against supervised baseline.

cially for some case where we only feel interested in the most similar retrieved objects. In addition, note that our model also achieve the highest generalised performance under FG-SBIR setting, showing that our model is able to capture semantic feature while also keeps the discriminativity.

6.2.4 Comparison against supervised baselines

We also compare with the supervised state-of-the-art baseline [3] trained in a supervised manor, *i.e.*, the target training data is visible under supervised setting for the baseline model updating. The results are shown in Tables 6-D. Though our model is not performing as well as the supervised baseline, it is still under an acceptable margin behind the competitor. However, the supervised baseline touches extra training data in the target categories while our method is not far behind by keep a well-generalised performance. The results reflect that the proposed model can be used out-of-the-box for any unseen domain.

Sketchy Dataset	Top 1	Top 10	mAP
w/o Logit-triplet	37.79%	85.13%	53.26%
w/o Matching Subnet	37.23%	85.50%	52.81%
Ours-full	**39.01%**	**86.44%**	**53.81%**

Table 6-E: Contributions of different components.

6.2.5 Ablation Study

There are two important components in the proposed DIMN: the matching subnet predicting the matching relations between the query sketches and gallery photos and the specifically designed logit-triplet loss built on the logit vector. To evaluate the contribution of each component, we compare our full model with the stripped-down version, which is obtained by removing the components. Note that by removing the matching subnet, we will also lose the logit-triplet loss as the latter is built on top of the matching subnet. Thus this ablated version is the same as the IDE baseline. Table 6-E shows that each component contributes the ReID performance.

6.2.6 Qualitative Result

Some qualitative results are shown in Fig. 6.3. In this figure, the left column represents the probe sketches randomly sampled from the testing dataset which spans 21 categories, while the gallery photos are searched with the top candidates returned as the retrieved result using DIMN. We use the green box to denote the ground-truth matched gallery images corresponding to the input probe image. From Fig. 6.3, it is clear to see that our method is able to achieve a generalised good performance on the unseen categories.

6.3 Summary

A generalisable fine-grained sketch-based image retrieval (FG-SBIR) approach was proposed to enable a FG-SBIR model to be deployed out-of-the-box for any new cate-

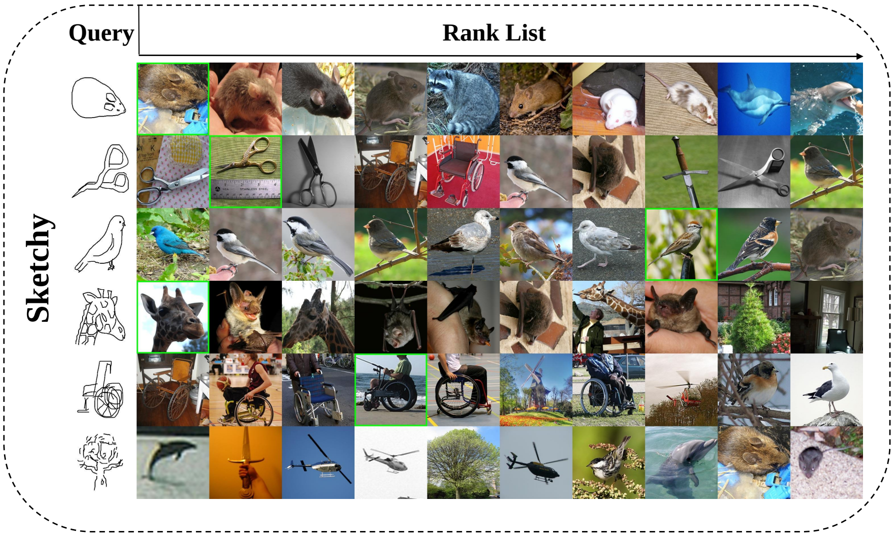


Figure 6.3: Retrieved result visualisation on the sketchy test set.

gory. Specifically, a novel deep FG-SBIR model termed Domain-Invariant Mapping Network(DIMN) was introduced. It has an encoding subnet to extract features from input sketches and photos and a mapping subnet that predicts a classifier weight vector from a single photo. The two subnets are trained end-to-end by using the mapping subnet as a hyper-network. The training follows a meta-learning pipeline to make the model domain invariant and generalisable to unseen categories. Extensive experiments on a newly defined large-scale benchmark validated the effectiveness of our DIMN. The experiments also showed that the generalisation in FG-SBIR is a very hard problem and existing FG-SBIR methods failed to achieve a generalised performance. However, given our promising results, and the practical value of a generalised FG-SBIR system, this is an important avenue for the FG-SBIR work. In the future work, we will also considering working on the generalisation learning for other sketch related tasks like sketch synthesis.

Chapter 7

Conclusions and Future Work

Sketch-related research has been advanced significantly with the help of collected sketch datasets and existing solid algorithms proposed based on those datasets, which have greatly exploited the promising potential of sketch modality. We contribute our idea and provide solutions for popular sketch-related tasks like fine-grained sketch-based image retrieval, sketch synthesis, and so on, as detailed in the previous chapters. In this chapter, we would like to conclude our contributions on sketch-related tasks, and share our views on future directions.

7.1 Conclusions

Sketch as a more and more popular modality involves attributes such as the informative, temporal, semantic and so on. The advantageous features of sketches have stimulated sketch-related research and shown great applicable value. In previous chapters we have discussed how to take advantage of the unique attributes of sketches to solve different challenging sketch-related tasks.

7.1.1 Fine-grained Sketch-based Image Retrieval

In fine-grained sketch-based image retrieval, one of our major contributions is the proposal of an attribute-driven deep multi-task framework which exploits semantic information from both sketch and photo modalities and narrows the domain gap between the two modalities. Our idea is based on the observation that existing FG-SBIR models depend excessively on the fine-grained representing ability of learned features and thus neglect semantic information. With our model, we are able to provide a more comprehensive and accurate ranking based on both high-level and middle-level information.

7.1.2 Sketch Synthesis

Sketch synthesis is another major direction for sketch-related research. Sketch synthesis research is very promising as it provides a deeper understanding of sketch modality and the sketching process, and also helps related tasks like FG-SBIR as it can generate synthetic data for training. We also contribute to the sketch synthesis community by contributing a deep photo-to-sketch synthesis framework, which can achieve satisfying results synthesising free-hand sketches. However, there are still some drawbacks to this framework which need to be addressed in future work. At first, the synthesis quality suffered in some instances where the trained model could not draw the fine-grained details accurately or drew the sketch in the wrong shape. In addition, the framework is not scalable as it needs to be applied to train different models for different categories. A universal neural sketcher is waiting to be designed which will be capable of drawing sketches according to given photos across multiple categories.

7.1.3 Generalisation for FG-SBIR

We also for the first time consider the generalisation problem for FG-SBIR tasks and design a specific model to search for a solution that produces acceptable performance for unseen categories. The proposed method can generalise a good performance for FG-SBIR on target unseen categories. However, the proposed method still has some drawbacks; for example, we only mimic the data sampling strategy from the meta-learning community, but we do not apply episode training to the model updating. In addition, there is a trade-off between the discriminability of the matching network and the precision of parameter updating, which also need to be reconsidered.

7.2 Future Works

7.2.1 Fine-grained Sketch-based Image Retrieval

Besides semantic information, there are also other clues omitted by existing deep FG-SBIR models. One important clue that needs to be exploited is part-level information. Part-level information is valuable to FG-SBIR tasks, as fine-grained features can also be represented in different parts of a given object in both sketch and photo modalities. One basic solution is based on the existing triplet-deep ranking model and integrates one of the part-level detection models to localise different part of query sketches and gallery photos. The designed model then needs to learn to rank in a part-level between sketches and photos. The final score will be a comprehensive ranking based on different parts as well as global features. The part-level localisation and ranking will release the constraint that the sketch and photo need to be aligned, and has the advantage of being able to describe which part distinguishes sketches and photos most by understanding part-level ranking scores.

As we indicate in another chapter, text descriptions can be very good compensation for sketches when doing FG-SBIR tasks. We thus propose a deep multi-modal retrieval network and also contribute a large multi-modal dataset with thousands of sketch-text-photo pairs. However, there is still much left to be done in this challenging task. In terms of datasets, a much larger multi-modal dataset is urgently needed, as the existing one has limited categories and instances, and current text descriptions are a bit beyond a common users' usage. In terms of the model, a more complicated framework is expected, which can learn to compensate between sketch and text modalities rather than simply aligning these two modalities as the current method [24] does.

7.2.2 Sketch Synthesis

Generative adversarial network (GAN) technology has been shown to greatly improve the image generation quality in style transfer tasks. Though photo-to-sketch synthesis is different from popular style transfer tasks which typically focus on image translation, it is still likely that generative adversarial training strategy will also help in sketch synthesis tasks. The simplest solution is to add an RNN discriminator and a related GAN loss, hoping that adversarial training can lead to a better RNN decoder to synthesise sketches.

We must also work urgently to solve the data issue in future works on sketch synthesis. QuickDraw has a very large scale, but the quality of the collected sketches is quite low and seeing some intact sketches is common because sketches are required to be finished in 20 seconds, thus it is hard for typical users to provide a good-quality drawing. The number of sketches in ShoeV2 and ChairV2 are still limited in order to train a better synthesis model. More categories are also required to form a more diverse dataset.

7.2.3 Generalisation for FG-SBIR

Based on the conclusions drawn for the proposed generalisation framework for FG-SBIR and its intrinsic limitations, further work can be dedicated to implementing the episode training for the proposed method, which will produce meta-learner generated optimised parameters for the generalised FG-SBIR task. At the same time, future work can also research a better way to seek both discriminativity and scalability. In the end, as indicated in the IBN-Net[121], instance normalisation is a better strategy to achieve the ability to generalise. Thus, one straightforward path ahead is to combine the instance normalisation strategy with our framework and generate a better performance based on this generalised FG-SBIR benchmark.

Our work also tries to let the research pay attention to the generalisation ability of existing FG-SBIR models, as the real-world application scenario strongly requires the generalisation ability of trained models. Moreover, we argue that the generalisation performance is also important to other sketch-related research. For example, though it is expensive and time-consuming to collect sketch and photo pairs to train a sketch synthesis model for new categories, with a generalised model we can expect an acceptable performance in synthesising sketches conditional on given photos coming from new categories without collecting and annotating new photos and sketches. To summarise, research to look for the generalised framework for other sketch-related works is promising.

7.2.4 Other Sketch Related Tasks

The advantages of sketches as a fine-grained and convenient input modality, are not only beneficial to the three mentioned areas like fine-grained sketch-based image retrieval, sketch-to-photo synthesis, and generalisation learning for FG-SBIR, but also very helpful for other tasks like sketch-to-video synthesis, sketch-aided 3D editing, and so on. For example, with a starting frame and the key actions/movements indicated by sketches, we are able to synthesis the whole video based on the photo and sketches. Also, for 3D editing, we can simply sketch the part we want to make changes, and the proposed model will automatically capture the modification guided by the sketch, and update the corresponding 3D model.

Appendix A

Author's publications

Conference Papers

1. **J. Song**, Y. Song, T. Xiang, T. Hospedales and X. Ruan, "Deep Multi-task Attribute-based Ranking for Fine-grained Sketch-based Image Retrieval," *British Machine Vision Conference (BMVC)* , York, UK, September, 2016.

2. **J. Song**, Y. Song, T. Xiang, and T. Hospedales, "Fine-Grained Image Retrieval: the Text/Sketch Input Dilemma," *British Machine Vision Conference (BMVC)*, London, UK, September, 2017.

3. **J. Song***, Qian Yu*, Y. Song, T. Xiang, and T. Hospedales, "Deep Spatial-Semantic Attention for Fine-Grained Sketch-Based Image Retrieval," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, October, 2017.

4. **J. Song**, K. Pang, Y. Song, T. Xiang, and T. Hospedales, "Learning to Sketch with Shortcut Cycle Consistency," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, June, 2018.

5. **J. Song**, Y. Yang, Y. Song, T. Xiang, and T. Hospedales, "Generalizable Person

Re-identification by Domain-Invariant Mapping Network," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, USA, June, 2019.

6. K. Li, K. Pang, **J. Song**, Y. Song, T. Xiang, and T. Hospedales, "Universal sketch perceptual grouping," *Proceedings of the the European Conference on Computer Vision (ECCV)*, Munich, GE, September, 2018.

7. K. Pang, D. Li, **J. Song**, Y. Song, T. Xiang, and T. Hospedales, "Deep Factorised Inverse-Sketching," *Proceedings of the the European Conference on Computer Vision (ECCV)*, Munich, GE, September, 2018.

8. A. Qi, **J. Song**, Y. Yang, Y. Song, T. Xiang, and T. Hospedales, "Fine-Grained Sketch-Based 3D Shape Retrieval with Cross-Modal View Attention," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Soul, Korea, September, 2019 (Submitted).

.

Journal Papers

1. Qian Yu*, **J. Song***, Y. Song, T. Xiang, and T. Hospedales, "Fine-Grained Instance-Level Sketch-Based Image Retrieval," *IEEE Transactions on Image Processing.* (Submitted)

Project Demos

1. **J. Song**, "Free-hand Sketch-Based Image Retrieval", `https://sketchx.eecs.qmul.ac.uk/`, 2016.

2. **J. Song**, "Free-hand Sketch Recognition", `https://sketchx1.eecs.qmul.ac.uk/`, 2017.

3. **J. Song**, "Photo-to-Sketch Synthesis", `https://sketchx2.eecs.qmul.ac.uk/`, 2018.

Bibliography

[1] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?" in *SIGGRAPH*, 2012.

[2] D. Ha and D. Eck, "A neural representation of sketch drawings," *ArXiv preprint arXiv:1704.03477*, 2017.

[3] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, and C. C. Loy, "Sketch me that shoe," in *CVPR*, 2016.

[4] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017.

[5] Y. Chen, S. Tu, Y. Yi, and L. Xu, "Sketch-pix2seq: a model to generate sketches of multiple categories," *arXiv preprint arXiv:1709.04121*, 2017.

[6] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.

[7] B. Paulson and T. Hammond, "Paleosketch: accurate primitive sketch recognition and beautification," in *ACMIUI*, 2008.

[8] T. Hammond and R. Davis, "Ladder, a sketching language for user interface developers," in *SIGGRAPH*, 2007.

[9] Y. Li, Y.-Z. Song, and S. Gong, "Sketch recognition by ensemble matching of

structured features." in *BMVC*, 2013.

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

[11] R. Hu, M. Barnard, and J. Collomosse, "Gradient field descriptor for sketch based retrieval and localization," in *ICIP*, 2010.

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

[13] J. M. Saavedra, "Sketch based image retrieval using a soft computation of the histogram of edge local orientations (s-helo)," in *ICIP*, 2014.

[14] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: Learning to retrieve badly drawn bunnies," in *SIGGRAPH*, 2016.

[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.

[16] K. Li, K. Pang, Y.-Z. Song, T. M. Hospedales, T. Xiang, and H. Zhang, "Synergistic instance-level subspace alignment for fine-grained sketch-based image retrieval," *TIP*, 2017.

[17] Y. Li, Y.-Z. Song, T. M. Hospedales, and S. Gong, "Free-hand sketch synthesis with deformable stroke models," *IJCV*, 2017.

[18] C. Chen, X. Tan, and K.-Y. K. Wong, "Face sketch synthesis with style transfer using pyramid column feature," in *WACV*, 2018.

[19] Y. Qi, Y.-Z. Song, T. Xiang, H. Zhang, T. Hospedales, Y. Li, and J. Guo, "Making better use of edges via perceptual grouping," in *CVPR*, 2015.

[20] K. Li, K. Pang, J. Song, Y.-Z. Song, T. Xiang, T. M. Hospedales, and H. Zhang, "Universal sketch perceptual grouping," in *ECCV*, 2018.

[21] U. R. Muhammad, Y. Yang, Y.-Z. Song, T. Xiang, T. M. Hospedales *et al.*, "Learning deep sketch abstraction," in *CVPR*, 2018.

[22] Y. Shen, L. Liu, F. Shen, and L. Shao, "Zero-shot sketch-image hashing," in *CVPR*, 2018.

[23] 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.

[24] J. Song, Y.-Z. Song, T. Xiang, and T. Hospedales, "Finegrained image retrieval: the text/sketch input dilemma," in *BMVC*, 2017.

[25] I. Berger, A. Shamir, M. Mahler, E. Carter, and J. Hodgins, "Style and abstraction in portrait sketching," *TOG*, 2013.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[27] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *ICCV*, 2017.

[28] J. Song, K. Pang, Y.-Z. Song, T. Xiang, and T. Hospedales, "Learning to sketch with shortcut cycle consistency," in *CVPR*, 2018.

[29] C. Hu, D. Li, Y.-Z. Song, and T. M. Hospedales, "Now you see me: Deep face hallucination for unviewed sketches," in *BMVC*, 2016.

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

[31] Q. Yu, Y. Yang, F. Liu, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Sketch-a-net: A deep neural network that beats humans," *IJCV*, 2017.

[32] A. Qi, Y.-Z. Song, and T. Xiang, "Semantic embedding for sketch-based 3d shape

retrieval." in *BMVC*, 2018.

[33] F. Wang, L. Kang, and Y. Li, "Sketch-based 3d shape retrieval using convolutional neural networks," in *CVPR*, 2015.

[34] X. Wang, X. Chen, and Z. Zha, "Sketchpointnet: A compact network for robust sketch recognition," in *ICIP*, 2018.

[35] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," *CVPR*, 2017.

[36] R. K. Sarvadevabhatla, J. Kundu *et al.*, "Enabling my robot to play pictionary: Recurrent neural networks for sketch recognition," in *ACMMM*, 2016.

[37] P. Xu, Y. Huang, T. Yuan, K. Pang, Y.-Z. Song, T. Xiang, T. M. Hospedales, Z. Ma, and J. Guo, "Sketchmate: Deep hashing for million-scale human sketch retrieval," in *cvpr*, 2018.

[38] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *ICCV*, 2003.

[39] L. Chen, D. Xu, I. W. Tsang, and J. Luo, "Tag-based web photo retrieval improved by batch mode re-tagging," in *CVPR*, 2010.

[40] W. Li, L. Duan, D. Xu, and I. W.-H. Tsang, "Text-based image retrieval using progressive multi-instance learning," in *ICCV*, 2011.

[41] L. Zheng, S. Wang, Z. Liu, and Q. Tian, "Packing and padding: Coupled multi-index for accurate image retrieval," in *CVPR*, 2014.

[42] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li, "Deep learning for content-based image retrieval: A comprehensive study," in *ACMMM*, 2014.

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

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

[45] T. Bui, L. Ribeiro, M. Ponti, and J. Collomosse, “Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network,” *CVIU*, 2017.

[46] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao, “Deep sketch hashing: Fast free-hand sketch-based image retrieval,” in *CVPR*, 2017.

[47] J. Zhang, F. Shen, L. Liu, F. Zhu, M. Yu, L. Shao, H. T. Shen, and L. Van Gool, “Generative domain-migration hashing for sketch-to-image retrieval,” in *ECCV*, 2018.

[48] F. Radenovic, G. Tolias, and O. Chum, “Deep shape matching,” in *ECCV*, 2018.

[49] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, “Cascade object detection with deformable part models,” in *CVPR*, 2010.

[50] P. Xu, Q. Yin, Y. Huang, Y.-Z. Song, Z. Ma, L. Wang, T. Xiang, W. B. Kleijn, and J. Guo, “Cross-modal subspace learning for fine-grained sketch-based image retrieval,” *Neurocomputing*, 2018.

[51] J. Song, Q. Yu, Y.-Z. Song, T. Xiang, and T. Hospedales, “Deep spatial-semantic attention for fine-grained sketch-based image retrieval,” in *ICCV*, 2017.

[52] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba, “Undoing the damage of dataset bias,” in *ECCV*, 2012.

[53] K. Muandet, D. Balduzzi, and B. Schölkopf, “Domain generalization via invariant feature representation,” in *ICML*, 2013.

[54] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, “Unified deep supervised domain adaptation and generalization,” in *ICCV*, 2017.

[55] S. Shankar, V. Piratla, S. Chakrabarti, S. Chaudhuri, P. Jyothi, and S. Sarawagi, "Generalizing across domains via cross-gradient training," in *ICLR*, 2018.

[56] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *AAAI*, 2017.

[57] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *arXiv preprint arXiv:1703.03400*, 2017.

[58] S. Thrun and L. Pratt, *Learning to learn.* Springer Science & Business Media, 2012.

[59] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *NIPS*, 2017.

[60] Q. Siyuan, L. Chenxi, S. Wei, and Y. Alan., "Few-shot image recognition by predicting parameters from activations," in *CVPR*, 2018.

[61] Q. Yu, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "SketchX! - Shoe/Chair fine-grained SBIR dataset," http://sketchx.eecs.qmul.ac.uk, 2017.

[62] C. Zou, Q. Yu, R. Du, H. Mo, Y.-Z. Song, T. Xiang, C. Gao, B. Chen, and H. Zhang, "Sketchyscene: Richly-annotated scene sketches," in *ECCV*, 2018.

[63] B. Li, Y. Lu, A. Godil, T. Schreck, M. Aono, H. Johan, J. M. Saavedra, and S. Tashiro, "Shrec13 track: large scale sketch-based 3d shape retrieval," in *3DOR*, 2013.

[64] B. Li, Y. Lu, C. Li, A. Godil, T. Schreck, M. Aono, M. Burtscher, H. Fu, T. Furuya, H. Johan *et al.*, "Shrec14 track: Extended large scale sketch-based 3d shape retrieval," in *3DOR*, 2014.

[65] B. Li, Y. Lu, F. Duan, S. Dong, Y. Fan, L. Qian, H. Laga, H. Li, Y. Li, P. Lui *et al.*, "Shrec'16 track: 3d sketch-based 3d shape retrieval," in *3DOR*, 2016.

[66] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "Sketch-based image retrieval: Benchmark and bag-of-features descriptors," *TVCG*, 2011.

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

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

[69] C. Wang, Z. Li, and L. Zhang, "Mindfinder: image search by interactive sketching and tagging," in *WWW*, 2010.

[70] Y. Cao, H. Wang, C. Wang, Z. Li, L. Zhang, and L. Zhang, "Mindfinder: interactive sketch-based image search on millions of images," in *ACMMM*, 2010.

[71] 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.

[72] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes," *TPAMI*, 2015.

[73] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo, "Multi-task deep visual-semantic embedding for video thumbnail selection," in *CVPR*, 2015.

[74] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," in *CVPR*, 2015.

[75] J. Moody, S. Hanson, A. Krogh, and J. A. Hertz, "A simple weight decay can improve generalization," in *NIPS*, 1995.

[76] F. Caron and A. Doucet, "Efficient bayesian inference for generalized bradley–terry models," *Journal of Computational and Graphical Statistics*, 2012.

[77] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *ICML*, 2009.

[78] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” in *ACMMM*, 2014.

[79] 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.

[80] L. Xie, J. Wang, B. Zhang, and Q. Tian, “Fine-grained image search,” *TMM*, 2015.

[81] E. Gavves, B. Fernando, C. G. M. Snoek, A. W. M. Smeulders, and T. Tuytelaars, “Fine-grained categorization by alignments,” in *ICCV*, 2013.

[82] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, “Part-based R-CNNs for fine-grained category detection,” in *ECCV*, 2014.

[83] Y. Jia, M. Salzmann, and T. Darrell, “Learning cross-modality similarity for multinomial data,” in *ICCV*, 2011.

[84] A. Mignon and F. Jurie, “Cmml: a new metric learning approach for cross modal matching,” in *ACCV*, 2012.

[85] T.-K. Kim, J. Kittler, and R. Cipolla, “Discriminative learning and recognition of image set classes using canonical correlations,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.

[86] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *CVPR*, 2015.

[87] M. Geng, Y. Wang, T. Xiang, and Y. Tian, “Deep transfer learning for person re-identification,” *arXiv preprint arXiv:1611.05244*, 2016.

[88] A. Graves, A.-r. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *ICASSP*, 2013.

[89] J. Wieting, M. Bansal, K. Gimpel, and K. Livescu, “Towards universal paraphrastic sentence embeddings,” *arXiv preprint arXiv:1511.08198*, 2015.

[90] L. Zitnick and P. Dollar, “Edge boxes: Locating object proposals from edges,” in *ECCV*, 2014.

[91] R. Řehůřek and P. Sojka, “Software framework for topic modelling with large corpora,” in *LRECW*, 2010.

[92] J. Huang, R. S. Feris, Q. Chen, and S. Yan, “Cross-domain image retrieval with a dual attribute-aware ranking network,” in *ICCV*, 2015.

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

[94] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, “A multi-view embedding space for modeling internet images, tags, and their semantics,” *IJCV*, 2014.

[95] W. Wang, R. Arora, K. Livescu, and J. Bilmes, “On deep multi-view representation learning,” in *ICML*, 2015.

[96] A. Graves, “Generating sequences with recurrent neural networks,” *arXiv preprint arXiv:1308.0850*, 2013.

[97] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, “Photo-realistic single image super-resolution using a generative adversarial network.” in *CVPR*, 2017.

[98] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.

[99] M. Mathieu, C. Couprie, and Y. LeCun, “Deep multi-scale video prediction beyond mean square error,” in *ICLR*, 2016.

[100] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen,

"Improved techniques for training gans," in *NIPS*, 2016.

[101] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T. Liu, and W.-Y. Ma, "Dual learning for machine translation," in *NIPS*, 2016.

[102] C. Li, H. Liu, C. Chen, Y. Pu, L. Chen, R. Henao, and L. Carin, "Towards understanding adversarial learning for joint distribution matching," in *NIPS*, 2017.

[103] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[104] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[105] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[106] L. Theis, A. v. d. Oord, and M. Bethge, "A note on the evaluation of generative models," in *ICLR*, 2016.

[107] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *ICML*, 2017.

[108] Y. Güçlütürk, U. Güçlü, R. van Lier, and M. A. van Gerven, "Convolutional sketch inversion," in *ECCV*, 2016.

[109] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays, "Scribbler: Controlling deep image synthesis with sketch and color," in *CVPR*, 2017.

[110] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," in *NIPS*, 2016.

[111] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *CoRR*, 2018.

[112] D. Ha, A. Dai, and Q. V. Le, "Hypernetworks," *arXiv preprint arXiv:1609.09106*, 2016.

[113] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018.

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

[115] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016.

[116] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[117] W. Deng, L. Zheng, G. Kang, Y. Yang, Q. Ye, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification," in *CVPR*, 2018.

[118] S. K. Yelamarthi, S. K. Reddy, A. Mishra, and A. Mittal, "A zero-shot framework for sketch based image retrieval." in *ECCV*, 2018.

[119] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *AISTAS*, 2010.

[120] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *TOMM*, 2017.

[121] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via ibn-net," in *ECCV*, 2018.